

医疗智能体

# 用户指南

文档版本 04  
发布日期 2025-01-22



版权所有 © 华为云计算技术有限公司 2025。保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

## 商标声明



HUAWEI和其他华为商标均为华为技术有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

## 注意

您购买的产品、服务或特性等应受华为云计算技术有限公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为云计算技术有限公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

# 目录

<b>1 用户指南（盘古辅助制药）</b> .....	<b>1</b>
1.1 欢迎使用盘古辅助制药平台.....	1
1.2 关键概念.....	1
1.3 准备工作.....	1
1.3.1 购买服务.....	2
1.3.2 购买资源.....	4
1.3.3 购买 CSS 资源.....	5
1.4 配额管理.....	8
1.5 系统设置.....	8
1.5.1 用户管理.....	8
1.5.1.1 创建子用户.....	8
1.5.1.2 导入用户.....	9
1.5.1.3 管理用户.....	11
1.5.1.4 用户权限介绍.....	12
1.5.2 资源中心.....	12
1.5.3 标签管理.....	13
1.5.4 个人设置.....	15
1.6 项目管理.....	15
1.6.1 项目简介.....	15
1.6.2 创建项目.....	17
1.6.3 项目成员和权限.....	18
1.7 功能模块.....	20
1.7.1 靶点发现.....	20
1.7.1.1 靶点口袋发现.....	20
1.7.1.2 靶点优化.....	22
1.7.2 苗头化合物发现.....	25
1.7.2.1 分子对接.....	25
1.7.2.2 分子属性预测.....	35
1.7.2.3 分子搜索.....	39
1.7.2.4 CPI 预测.....	41
1.7.2.5 分子生成.....	44
1.7.3 先导化合物优化.....	52
1.7.3.1 分子优化.....	52

1.7.3.2 靶点口袋分子设计.....	60
1.7.3.3 自由能微扰.....	66
1.7.3.4 合成路径规划.....	70
1.7.4 通用工具.....	72
1.7.4.1 MOL 3D Viewer.....	72
1.7.4.2 MOL Editor.....	72
1.7.4.3 聚类分析.....	73
1.8 AI 模型.....	75
1.9 自定义数据库.....	76
1.10 数据管理.....	78
1.10.1 数据管理简介.....	78
1.10.2 添加数据.....	79
1.10.3 数据管理常用操作.....	83
1.11 作业管理.....	85
1.11.1 作业管理简介.....	85
1.11.2 创建作业.....	86
1.11.3 管理作业.....	87
1.11.4 查看执行结果.....	87
1.12 收藏夹.....	89
1.13 相关参数.....	91
<b>2 用户指南（基因平台）.....</b>	<b>101</b>
2.1 欢迎使用医疗智能体服务.....	101
2.2 关键概念.....	102
2.3 准备工作.....	103
2.3.1 购买服务.....	103
2.3.2 购买存储套餐包.....	104
2.3.3 购买系统资源.....	106
2.3.4 获取认证信息.....	116
2.4 用户管理.....	117
2.4.1 用户管理介绍.....	117
2.4.2 创建子用户.....	118
2.4.3 导入用户.....	119
2.4.4 登录平台方式.....	121
2.4.5 管理用户.....	122
2.5 配额管理.....	123
2.6 系统设置.....	124
2.6.1 安全设置.....	124
2.6.2 商标设置.....	124
2.6.3 作业消息设置.....	125
2.6.4 消息中心.....	125
2.6.5 邮箱设置.....	129
2.6.6 系统标签管理.....	131

2.6.7 个人资源配额.....	133
2.6.8 数据归档配置.....	134
2.7 项目管理.....	135
2.7.1 项目简介.....	135
2.7.2 创建项目.....	137
2.7.3 项目成员和权限.....	138
2.7.4 项目管理常用操作.....	142
2.8 数据管理.....	145
2.8.1 数据管理简介.....	145
2.8.2 添加数据.....	145
2.8.3 发布数据.....	148
2.8.4 数据管理常用操作.....	150
2.8.5 归档数据.....	154
2.8.6 数据控制与数据审计.....	155
2.8.7 命令行工具概述.....	156
2.9 数据库管理.....	164
2.9.1 数据库简介.....	164
2.9.2 创建数据库模板.....	165
2.9.3 导入数据库模板.....	166
2.9.4 创建数据库.....	167
2.9.5 引用数据库.....	169
2.9.6 管理数据库.....	169
2.10 镜像管理.....	170
2.10.1 镜像管理简介.....	170
2.10.2 导入或上传镜像.....	171
2.10.3 安装容器引擎.....	172
2.10.4 获取镜像.....	172
2.10.5 制作 Docker 镜像.....	173
2.10.6 上传镜像到 SWR 镜像仓库.....	178
2.10.7 发布镜像.....	179
2.11 工具管理.....	180
2.11.1 工具管理简介.....	180
2.11.2 新建应用.....	181
2.11.3 导入应用.....	183
2.11.4 上传应用.....	183
2.11.5 发布应用.....	185
2.11.6 编辑应用.....	186
2.11.7 创建 FastQC 应用样例.....	186
2.11.8 新建流程.....	189
2.11.9 流程设计器.....	193
2.11.10 导入流程.....	194
2.11.11 上传流程.....	194

2.11.12 发布流程.....	196
2.11.13 添加分类标签.....	197
2.11.14 下载应用或流程.....	197
2.12 作业管理.....	198
2.12.1 分析作业管理简介.....	198
2.12.2 创建分析作业.....	198
2.12.3 创建自动作业.....	206
2.12.4 查看执行结果.....	209
2.12.5 作业执行失败排查思路.....	212
2.13 Nextflow.....	220
2.13.1 简介.....	220
2.13.2 安装/卸载 Nextflow.....	220
2.13.3 新建流程.....	221
2.13.4 管理流程.....	223
2.13.5 新建作业.....	224
2.13.6 管理作业.....	225
2.13.7 约束限制.....	226
2.13.8 作业执行失败排查思路.....	231
2.14 资产市场.....	232
2.14.1 资产市场简介.....	232
2.14.2 使用 Variant Calling Based On NGS 流程.....	236
2.14.3 使用 Docking Summary 流程.....	239
2.14.4 使用 RNA-Seq Analysis Based on STAR 流程.....	242
2.14.5 使用 AutoGenome 镜像.....	246
2.14.6 genomeagent 镜像.....	248
2.14.6.1 使用 genomeagent 镜像.....	248
2.14.6.2 genomeagent 镜像功能点介绍.....	254
2.15 开发环境 ( Notebook ) .....	257
2.15.1 Notebook 简介.....	257
2.15.2 创建 Notebook.....	258
2.15.3 Notebook 常用操作.....	259
2.15.4 数据的上传和下载.....	262
2.15.5 自定义镜像创建 Notebook 样例.....	263
2.15.6 JupyterLab 简介及常用操作.....	266
2.15.7 url 获取方法.....	269
2.15.8 Notebook 安装 Conda 指导.....	270
2.16 大规模药物虚拟筛选.....	273
2.16.1 “神农项目”简介.....	273
2.16.2 药物虚拟筛选.....	274

# 1 用户指南（盘古辅助制药）

## 1.1 欢迎使用盘古辅助制药平台

盘古辅助制药平台是以盘古药物大模型为基础打造的一站式药研平台，助力药物研发效率提升60%+。平台提供靶点发现，苗头化合物发现，先导化合物优化全流程药研所需功能。同时基于云原生的软硬件一体化加速，大大提升虚拟筛选和分子动力学模拟计算效率。

全平台无需软硬件安装调试成本，开箱即用，随时可用。全平台包括盘古辅助制药赋能药物研发的全链条任务，将新药研发从“马拉松”迈向“加速跑”。

## 1.2 关键概念

### 盘古药物分子大模型

盘古药物分子大模型是基于华为与中科院上海药物所共同研发、专门面向药物研发领域推出的预训练大模型，旨在帮助医药公司开启AI辅助药物研发的新模式。盘古药物分子大模型学习了17亿个药物分子的化学结构，模型参数上亿，是目前最大的小分子药物模型。华为盘古药物分子大模型在分子生成、属性预测、生物活性预测和分子优化等20多个药物发现任务上均达到性能最优。

### SPONGE

分子模拟是指利用计算机以原子水平的分子模型来模拟分子结构与行为，进而模拟分子体系的各种物理、化学性质的方法。它是在实验基础上，通过基本原理，构筑起一套模型和算法，从而计算出合理的分子结构与分子行为。SPONGE（Simulation Package toward Next Generation）是由北京大学高毅勤教授课题组与华为团队联合开发的新一代分子动力学模拟程序，具有高性能、模块化等特性，是一个完全自主研发的分子模拟软件库。基于高毅勤教授课题组和华为团队的技术支持，已经实现自由能微扰加速10倍以上。测试

## 1.3 准备工作

## 1.3.1 购买服务

### 前提条件

- **注册华为账号并开通华为云**，并进行**账号实名认证**。购买服务时，账号不能处于欠费或冻结状态。
- 购买服务时，需使用华为主账号购买。如果不确定账号类型，可登录**账号中心**查看所属账号是主账号还是子账号（IAM账号）。  
在账号中心左侧导航栏中有“我的主账号”页签，则说明是子账号；没有“我的主账号”页签，则说明是主账号。

图 1-1 查看账号信息



- 联系技术支持开通白名单。  
需要提供账号名，账号ID，项目ID等信息。相关信息获取方式如下：
  - 使用华为云账号登录后，在右上角账号下选择“我的凭证”。



- 在API凭证中可以查看相关信息。



## 购买服务

盘古辅助制药平台提供包年包月、按需两种计费方式，您可以按需求进行选择。关于计费模式的详细介绍，请参见[计费说明](#)。

1. 登录华为云[管理控制台](#)，在左侧服务列表中选择“人工智能 > 医疗智能体 EIHealth”。

图 1-2 服务列表



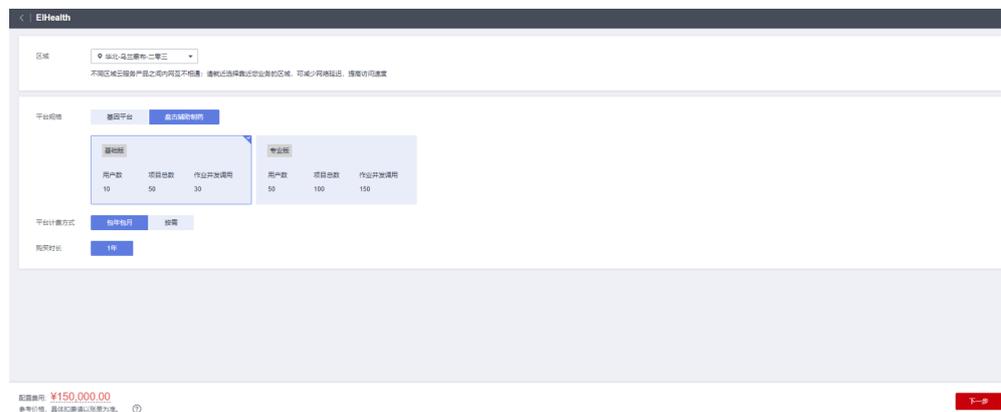
2. 进入医疗智能体“总览”页面，单击“购买平台”。

图 1-3 购买平台



3. 购买盘古辅助制药平台。

图 1-4 购买盘古辅助制药



- 区域：请就近选择靠近您业务的区域，可减少网络延迟，提高访问速度。不同区域云服务产品之间内网互不相通。
  - 平台规格：选择“盘古辅助制药平台”基础版或者专业版。
  - 平台计费方式：选择“包年包月”或“按需”计费。
  - 购买时长：如果选择的“包年包月”计费，目前只支持购买一年；如果选择的“按需”计费，无需选择购买时长。
- 勾选自动续费后，系统将在产品到期前自动续费，无需用户再手动操作。按月购买，自动续费周期为一个月。
4. 单击“下一步”，确认订单详情。
  5. 单击“立即购买”，并支付订单，完成服务购买。
- 购买成功后，需要大约30分钟时间完成服务自动部署，请耐心等待。部署完成后，您可以使用注册的华为账号使用医疗智能体平台。

### 1.3.2 购买资源

盘古辅助制药平台购买完成后，单击“进入平台”，在平台右上角单击用户名，选择“资源中心”。

您可以根据实际需求选择购买功能调用套餐包和存储套餐包。在后续使用过程中，也可根据使用情况随时购买资源。

### 资源中心

在“资源中心>功能调用”中可以查看功能调用套餐包的使用情况。

图 1-5 功能调用套餐包



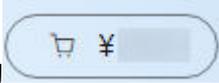
在“资源中心>存储资源”中可以查看存储使用量等信息。

图 1-6 存储资源信息



## 购买功能调用套餐包

1. 在“资源中心>功能调用”页面，购买功能调用套餐包。
2. 根据需求，选择套餐包类型。目前支持200次/年、600次/年、2500次/年。

3. 单击套餐包右下角的 ，进行支付。  
套餐包购买成功后，不支持退订，不能抵扣已经产生的用量。

## 购买存储套餐包

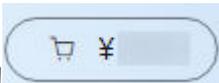
1. 在“资源中心>存储资源”页面，购买存储套餐包。
2. 根据需求，选择套餐包类型，单击套餐包右下角的 ，进行支付。  
或者单击右上角“更多”，进入购买存储套餐包页面进行购买。

表 1-1 参数说明

参数	说明
资源包类型	支持标准存储单AZ包。 标准存储单AZ包适合数据分享、内容分享、热点对象应用场景。
规格	套餐包支持的购买规格。
购买数量	购买套餐包的数量。取值范围：1-50。
购买时长	购买套餐包的时长。最短为一个月，最长为一年。
生效时间	选择购买套餐包后的生效时间。目前支持支付完成后立即生效和指定生效时间两种方式。 如果支付时间晚于指定生效时间，资源包将在支付后立即生效。

套餐包购买成功后，不支持退订，不能抵扣已经产生的用量。

### 1.3.3 购买 CSS 资源

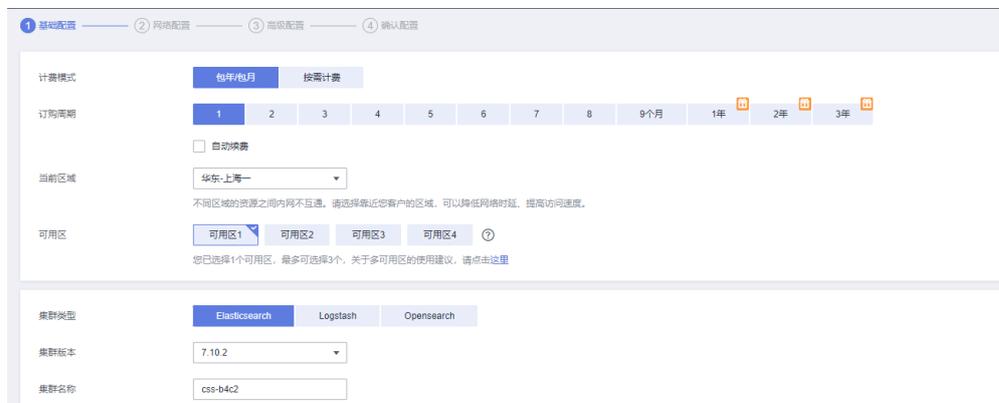
1. 登录[CSS云搜索](#)服务。
2. 单击“创建集群”。

图 1-7 创建集群



3. 基础配置：选择“计费模式”、“当前区域”。  
集群类型选择“Elasticsearch”，输入集群名称。

图 1-8 基础配置



节点规格：参考自己的数据库大小。自定义数据库所需要的的CSS节点规格大于“4 vCPUs | 8GB”，不支持“4 vCPUs | 8GB”。

节点存储：建议选择“超高I/O”。

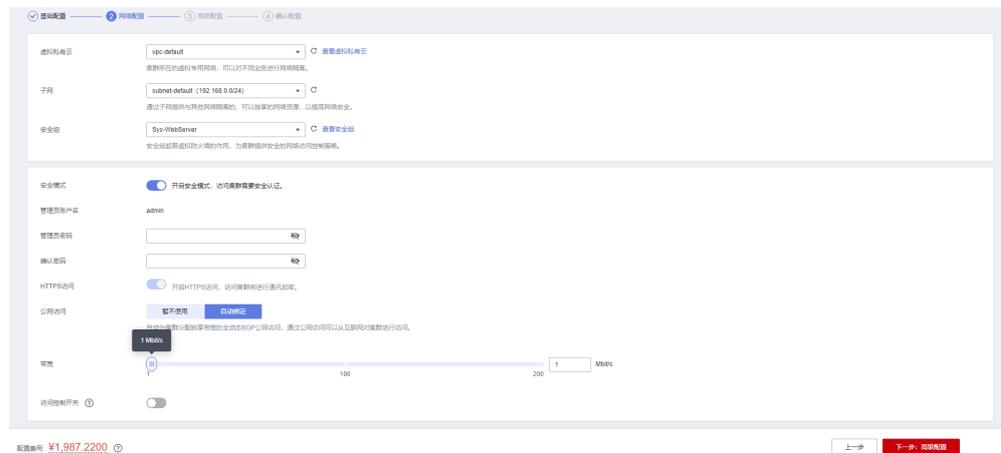
单击下一步“网络配置”。

图 1-9 节点



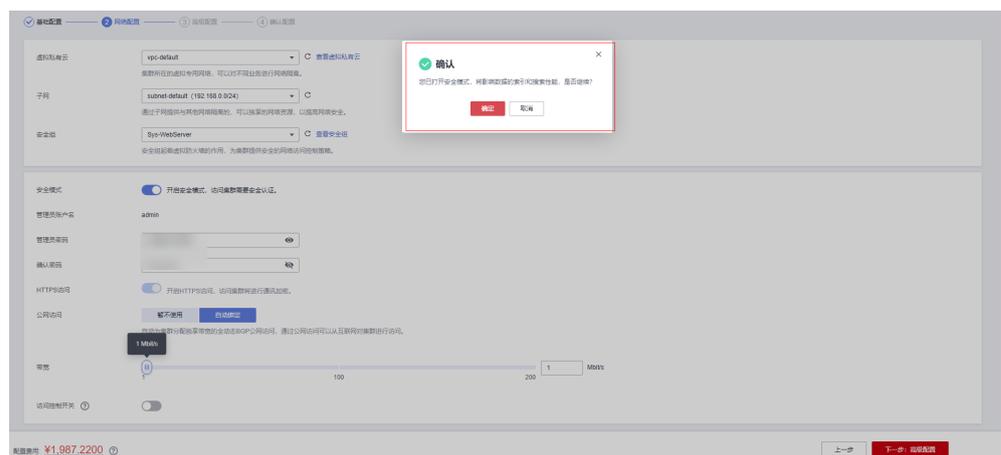
4. 网络配置：输入管理员密码，公网访问选择“自动绑定”，单击“下一步高级配置”。

图 1-10 网络配置



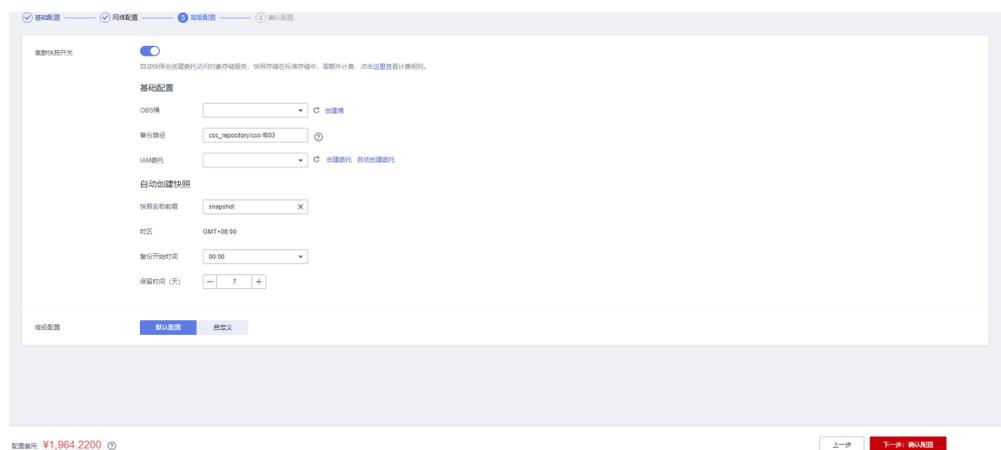
单击“下一步：高级配置”，在弹窗里单击“确认”。

图 1-11 确认



5. 高级配置。

图 1-12 高级配置



6. 确认配置，无误后单击“立即创建”，即可完成购买。

图 1-13 确认配置

基础信息	当前版本: 7.10.2 当前类型: elasticsearch 订购日期: 11月 支付方式: 否 企业代码: default 计费模式: 按年收费			
硬件配置	节点规格: emr-spec-3u15g   8 vCPU   16 GB 存储配置: 30 GB 数据IO			¥1,332,3400
网络配置	所属区域: 华东-上海一 可用区: cn-east-3a 虚拟私有云: vpc-eeat8con 子网: subnet-1902   192.168.0.0/24 安全组: sg-eeat8con 版本描述: 通用 HTTP代理: 关闭 公网访问: 独享带宽   主网带宽 1 Mbps			¥23,0000
资源许可策略	未开启			
集群配置	集群名称: 开发 ODC账号: elasticsearch-pydata-cn-east-3 角色描述: cn_apollo/tao-4-4-2 IAM策略: cn_apollo_iam-4-4-2 快照策略: snapshot HDFS: GMT+08:00 默认时区: GMT+08:00 保留时间 (天): 7			
Kibana公网访问	未开启			
标签	-			

总价格: ¥1,332,3400

上一步 下一步

## 1.4 配额管理

盘古辅助制药平台分为基础版和专业版两个版本。

表 1-2 基础版与专业版区别

功能配额	专业版	基础版
用户总数	50	10
项目总数	100	50
单用户创建项目数	20	20
作业并发数	150	30
模型训练&模型管理	支持	不支持

## 1.5 系统设置

### 1.5.1 用户管理

#### 1.5.1.1 创建子用户

1. 使用管理员账户登录盘古辅助制药平台。
2. 在右上角用户名中选择“用户管理”。
3. 在用户管理页面，单击左上角“创建用户”，进入“创建用户”页面。
4. 填写用户信息，分配子用户权限。

5. 创建完成后，单击“确认”。

### 1.5.1.2 导入用户

盘古辅助制药平台支持把IAM子用户导入至平台子用户。通过导入已有子用户，增加用户使用的便捷度，方便用户维护账号。

1. 使用管理员账户登录盘古辅助制药平台。
2. 在右上角用户名中选择“用户管理”。
3. 在用户管理页面，单击“导入用户”，进入“导入用户”页面。
4. 在导入用户页面，可以选择“用户”或者“用户组”进行导入。

图 1-14 导入用户

## 说明

- 导入的IAM子用户需要具有管理控制台访问方式。
  - 导入用户时不能超出配额。如果超出配额，进行配额调整后，5分钟后生效。
  - 以用户组的方式导入时，若超出配额的部分会导入失败。
  - 以用户组的方式导入时，用户组里已经导入到平台的用户，不算统计个数。例如，用户组A里50个用户，10个已经导入平台，那么统计时，只会显示已选择40个用户。
  - IAM平台限制一个IAM用户不能加入超过10个用户组，而导入时还会再加入医疗用户组。如果IAM用户在导入平台之前就已经加入了10个用户组，则导入的时候会失败。
  - 用户数统计时会去重。例如，用户组A有50个用户，用户组B里有50个用，两者之间有10个用户重复，那么同时选择用户组A和B，统计时显示已选择90个用户。
  - 如果导入后用户状态显示异常，需联系技术支持处理。
5. 单击“下一步”，设置角色。设置是否为“系统管理员”。
  6. 配置完成后，单击“确定”。
  7. 等待导入成功后，单击“关闭”。可以在用户管理页面查看导入成功的用户信息。

导入的用户，不支持删除，只支持移除，移除后不影响该用户操作其他服务。

图 1-15 查看导入用户



## 说明

医疗平台用户会在IAM中赋予以下细粒度权限，若该用户加入的其他IAM用户组有对应的deny权限，则会影响平台部分功能使用。

```
{
  "Version": "1.1",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "obs:bucket:GetBucketLocation",
        "obs:bucket:ListBucketMultipartUploads",
        "obs:object:GetObjectVersion",
        "obs:object:GetObject",
        "obs:object:GetObjectVersionAcl",
        "obs:object:DeleteObject",
        "obs:object:ListMultipartUploadParts",
        "obs:bucket:HeadBucket",
        "obs:object:AbortMultipartUpload",
        "obs:object:DeleteObjectVersion",
        "obs:object:GetObjectAcl",
        "obs:bucket:ListAllMyBuckets",
        "obs:bucket:ListBucket",
        "obs:object:PutObject"
      ]
    }
  ]
}
```

### 1.5.1.3 管理用户

#### 删除子用户

管理员用户包含子管理员，均具备删除权限。但子管理员不可将管理员删除。删除用户只能删除来源为本平台的用户，如果来源为IAM，则只支持移除。

在用户的操作列，单击“删除”，删除对应的用户。

图 1-16 删除子用户



#### 移除子用户

导入的用户，不支持删除，只支持移除。移除后不影响该用户操作其他服务。执行移除用户操作时，只有在该用户名下没有项目时，才可以被移除。在用户的操作列，单击“移除”，移除对应的用户。

图 1-17 移除子用户



### 安全设置

在用户管理页面，可以重置用户的邮箱、手机号以及系统管理员权限的更改。在用户操作列，单击“编辑”，重置用户的相关信息。

最终租户可以编辑子用户角色、手机号、邮箱。

系统管理员只能编辑角色，非系统管理员无编辑权限。

图 1-18 重置用户信息



## 重置密码

最终租户在用户管理页面，可以重置用户的密码。在用户操作列，单击“重置密码”，重置用户的密码信息。

图 1-19 重置密码



### 1.5.1.4 用户权限介绍

表 1-3 用户权限说明

用户类别	权限
最终租户	创建用户、导入用户、删除用户、编辑角色、邮箱和手机号、重置密码、移除用户。
系统管理员	创建用户、导入用户、删除用户、编辑角色、移除用户。
非系统管理员	无相关编辑权限。

## 1.5.2 资源中心

在资源中心可以查看功能调用套餐包和存储套餐包的使用情况、购买记录以及购买套餐包，也可以绑定CSS资源详细的购买步骤可以参考[购买资源](#)，此功能只支持专业版。

### 绑定 CSS 资源

使用自定义数据库之前需要先购买和绑定CSS资源，购买步骤参考[购买资源](#)，绑定具体操作如下：

1. 在盘古辅助制药平台单击右上角“账号名>资源中心”进入CSS资源界面。单击“绑定”进入绑定界面，选择要绑定的CSS集群名称、填写管理员账户名和管理员密码。单击“测试链接”验证用户名或密码无误后单击“确定”，即可完成绑定（可选项）。您也可以单击操作列的“解除绑定”来解除当前的绑定。

图 1-20 绑定 CSS 资源入口



图 1-21 绑定 CSS 资源

**绑定CSS资源**

\* CSS集群名称

请选择

若CSS集群显示“不可绑定”，请查看 [CSS常见问题指导](#)

\* 管理员账户名

admin

\* 管理员密码

请输入

测试连接

取消 确定

### 1.5.3 标签管理

在平台右上角用户名中选择“标签管理”。

系统标签用于设置作业分类，可以在系统标签管理中添加、删除标签。

#### 添加标签

1. 在“标签管理”页面，单击“添加”。
2. 在“添加标签”弹窗中，输入标签名称、标签描述。

标签名称长度为1~32，支持中文、字母、数字、空格、下划线和中划线，且不能以空格开头或者结尾。

图 1-22 添加标签



3. 单击“确定”，标签添加完成。系统最多可添加100个标签。

## 删除标签

标签删除后不可恢复，请谨慎操作。

1. 在“标签管理”页面，单击操作列的“删除”。
2. 在“删除标签”弹窗中，单击“确定”。

## 批量删除标签

您可以一次选中多个标签，单击页面左上角的删除按钮即可。

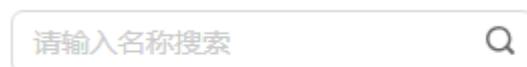
图 1-23 批量删除标签



## 查找标签

在搜索框中输入标签名称关键字，单击  按钮进行查找。

图 1-24 查找标签



## 1.5.4 个人设置

在平台右上角用户名中选择“个人设置”，可以修改邮箱、手机号和密码。最终租户仅可修改邮箱和手机号，子用户可以修改可以修改邮箱、手机号和密码。

图 1-25 修改个人信息



## 1.6 项目管理

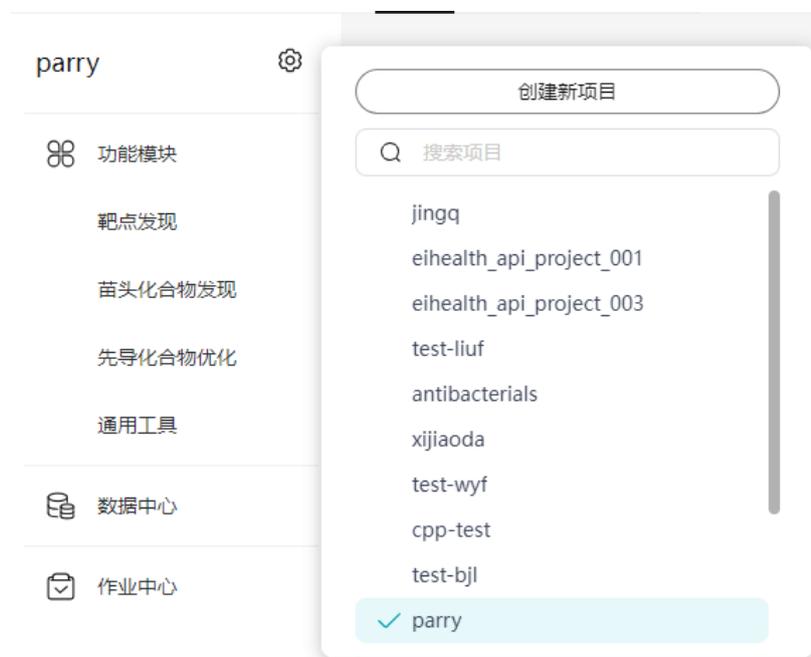
### 1.6.1 项目简介

项目是盘古辅助制药平台的一个工作空间，可以在项目中存储数据和创建作业。也可以将团队成员引入到项目中，并通过设置成员角色实现项目权限的划分。

您可以创建项目，并向其中上传数据、创建作业。

在“项目列表”中，展示了当前用户有权限访问的项目。

图 1-26 项目列表



### 查看项目信息

项目信息中提供了项目设置，成员管理，数据审计，您可以通过如下方式查看。

在平台左上角选择项目名称，单击  ，进入项目设置页面。

图 1-27 查看项目信息



- 项目设置

在项目设置页面，可以查看项目名称，该项目的数据存储OBS桶名称。所有者可以修改项目描述，转移项目，删除项目。管理员可以修改项目描述，退出项目。操作员可以退出项目。

删除项目后数据、作业会同步删除，删除项目不可逆，请谨慎操作。

- 成员管理

在成员管理页面可以添加或者移除访问该项目的成员。

图 1-28 成员管理



- 数据审计

平台通过云审计服务（CTS）提供操作记录的收集、存储和查询，审计操作可以设置导出用户的读操作和写操作（默认）和仅写操作（可选），保存周期可以设置需要保存的审计日志的保存周期，下载按钮可以下载最近7天内最新的1万条数据审计日志，查看按钮可以查看及下载保存的审计日志。

通过审计日志可用于支撑安全分析、合规审计、资源跟踪和问题定位等常见应用场景。您可以在项目的“数据审计”页面下载最近7天的数据审计日志，其他操作的审计日志请登录云审计服务控制台查看。

图 1-29 数据审计

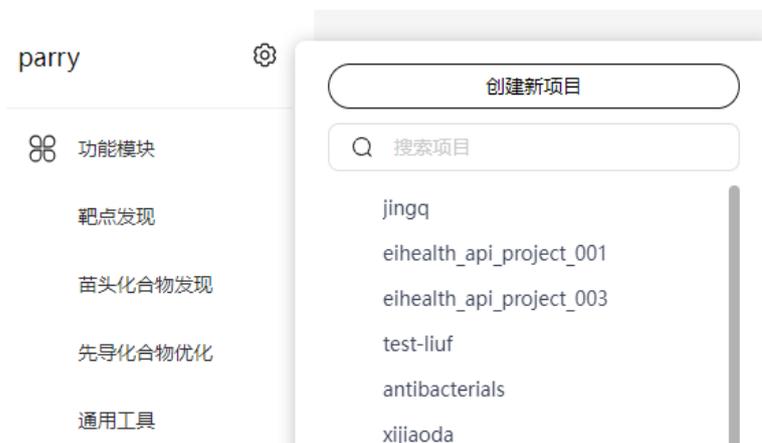


## 1.6.2 创建项目

您可以创建一个新的项目。

1. 在平台左上角单击项目名称，选择“创建新项目”。

图 1-30 创建新项目



2. 配置项目信息。

表 1-4 参数说明

参数	说明
项目名称	项目名称。取值范围3~45个字符，包含数字、小写字母、下划线、中划线，且只能以小写字母数字开头结尾。 项目创建成功后，不支持修改名称。
描述	设置项目描述。

图 1-31 创建项目

3. 单击“确认”，创建一个新的项目。

#### 📖 说明

- 项目的创建者默认拥有项目的完整权限，同时项目可以分享给其他用户，并限定其他用户的访问权限。项目角色为项目粒度权限控制，同一用户在不同的项目上可能拥有不同的角色。
- 创建的项目配额请参见[配额管理](#)进行查询。详细添加项目成员并分配角色的方法请参见[添加项目成员](#)。

## 1.6.3 项目成员和权限

盘古辅助制药平台以项目为粒度对数据、作业进行分组。以使用户通过项目进行资源的访问、共享和协作。

- [添加项目成员](#)
- [移除、修改项目成员](#)
- [成员角色和权限](#)

### 添加项目成员

#### 前提条件

- 平台管理员首先通过“用户管理”功能添加平台用户，才能将该用户添加至项目中。  
创建用户的详细方法请参见[创建平台用户](#)。
- 存在一个创建好的项目。
  1. 单击项目名称后面  ，进入项目设置页。
  2. 选择“成员管理”，单击“添加成员”。

图 1-32 添加成员



3. 选择已添加至平台的成员名称及权限。

图 1-33 选择成员



4. 单击“确定”，将用户添加至项目中。

## 移除、修改项目成员

1. 单击项目名称后面  ，进入项目设置页。
2. 选择“成员管理”，在“权限”列重新设置项目成员角色。  
详细项目成员角色和权限介绍请参见[成员角色和权限](#)。

图 1-34 移除、修改项目成员



- 选中成员，单击操作列“移除”，将成员从项目中移除。

## 成员角色和权限

在项目中，成员可以划分为5种不同的角色。详细角色介绍请参考[表 项目成员角色](#)。

表 1-5 项目成员角色

角色	说明
创建者	拥有项目全部操作：成员导入、转移项目、删除项目等。
管理员	除了不能转移项目、删除项目，其他都可以比如导入成员。
操作员	不能转移项目、删除项目、导入成员，其他都可以。

### 说明

- 项目列表中只显示您所在的项目。
- 一个用户可以是多个项目的成员。

## 1.7 功能模块

### 1.7.1 靶点发现

#### 1.7.1.1 靶点口袋发现

靶点口袋发现实现全新元动力学算法，高效遍历柔性蛋白表面，找到潜在口袋位置。

- 单击“靶点口袋发现”功能卡片，进入配置页面。
- 在配置页面，选择靶点文件、探针文件，配置相关参数信息。
  - 靶点文件：支持PDB格式文件。
  - 探针文件：探针分子建议原子数不超过400，分子量过大可以通过增大探针半径降低失败率，但无法保证一定成功。支持SDF、MOL2、PDB格式文件。若文件中含有多个配体，默认只处理第一个。
  - 时间步长：默认值为0.002ps，输入范围：0.001<dt<=0.002，单位：ps。分子动力学模拟的步长，建议不超过0.002ps，步长越大，越难收敛。
  - 最小化步数：默认值为20000，输入范围：5000-50000。
  - 预平衡时长：默认值为100ps，输入范围：20-200，单位：ps。时长增加，作业运行时间延长。
  - 口袋发现时长：默认值为50ns，输入范围：20-50。
  - 表面原子离散点数量：每个表面原子离散点数量。默认值为20，输入范围：10-50。
  - 探针半径：控制产生的离散点到表面原子的距离。默认值为1.4，输入范围：1.4-5，单位：Å。

- 名称：可修改，修改后左上角也同步修改。长度为5~64个字符；仅可以使用字母、数字、下划线“\_”、中划线“-”和空格；首位只能以数字或字母开头。
  - 标签：设置任务标签。
  - 功能调用消耗：运行一次功能会消耗一次。
3. 单击“提交”。
- 提交成功后，可以在“作业中心”查看执行结果。

图 1-35 设置靶点口袋发现文件



4. 查看靶点口袋发现结果。
- 单击作业名称，在输出结果页面，可以单击“查看轨迹”，下载运动轨迹。可以单击”下游分析”，将发现的口袋作为对接口袋进行分子对接。也可以在右边口袋列表中查看口袋信息。

单击口袋列表的  ，可以收藏靶点口袋发现结果，收藏后可在收藏夹页进行查看。

图 1-36 查看靶点口袋发现结果

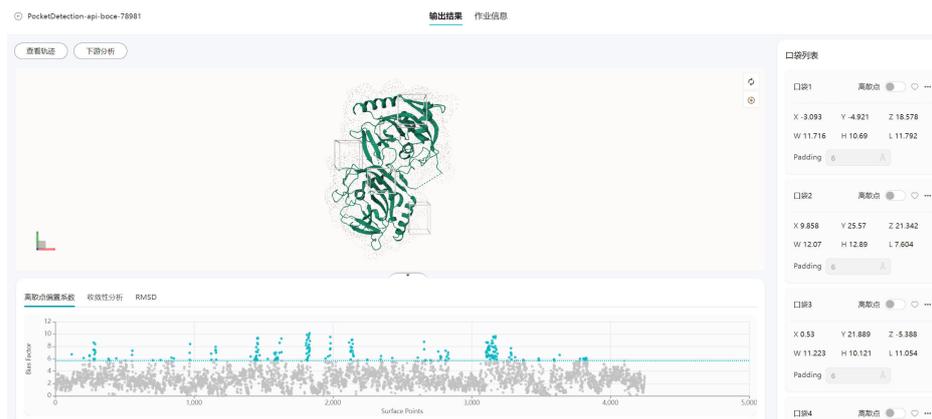
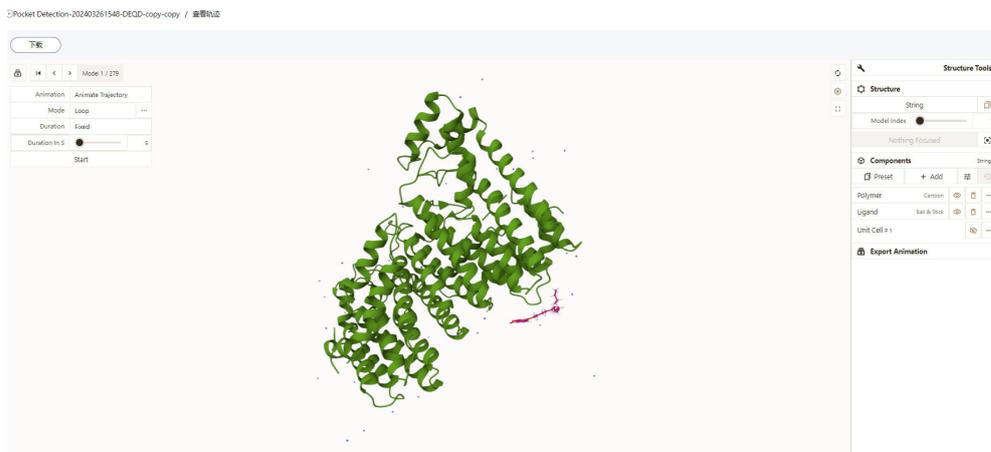


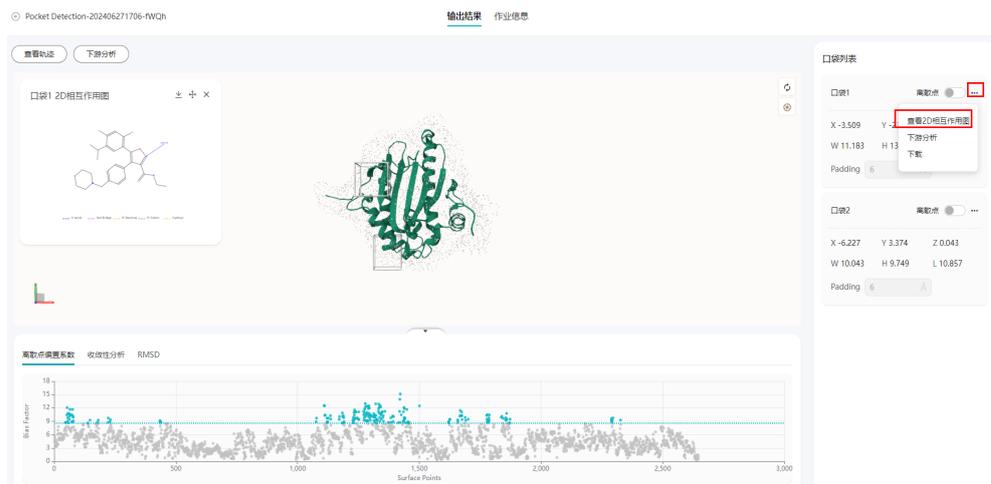
图 1-37 查看靶点口袋发现运动轨迹



5. 查看2D相互作用图。

在“作业中心”单击作业名称，在输出结果页面右侧口袋列表，单击口袋后面的图标 ，然后单击查看2D相互作用图即可。

图 1-38 查看 2D 相互作用图



6. 查看作业信息图。

在“作业中心”单击作业名称，单击“作业信息”页面。可以查看作业的参数和运行信息。

图 1-39 查看作业信息图

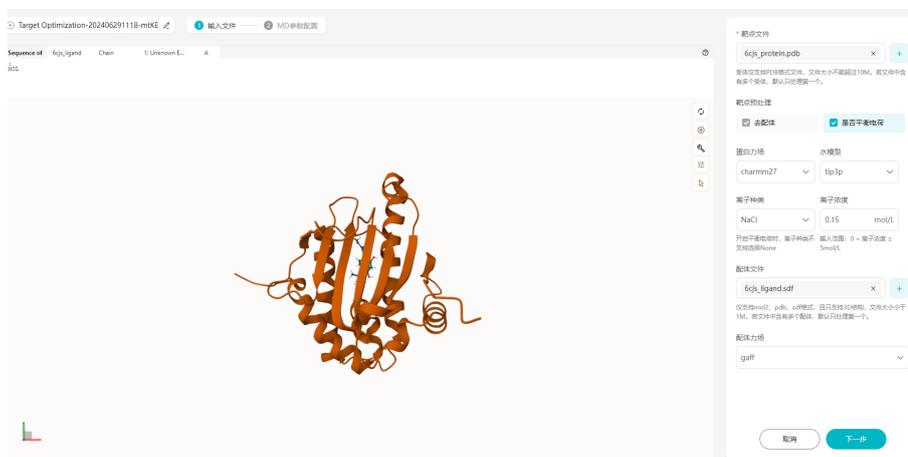


### 1.7.1.2 靶点优化

靶点优化基于分子动力学模拟和结构聚类，实现靶点结构优化

1. 单击“靶点优化”功能卡片，进入配置页面。
2. 配置靶点文件和相关参数信息。
  - 靶点文件：支持PDB格式文件，文件大小不能超过10M。若文件中含有多个受体，默认只处理第一个。
  - 靶点预处理：
    - 去配体：提交任务时系统会自动删除配体。
    - 是否平衡电荷：是否加入离子，使体系电荷守恒，默认是平衡电荷。
  - 蛋白力场：蛋白质力场支持amber03, amber94, amber96, amber99, amber99sb, amber99sb-ildn, amberGS, charmm27, oplsa, gromos43a1, gromos43a2, gromos45a3, gromos53a5, gromos53a6, gromos54a7力场，默认是amber99sb。
  - 水模型：水模型支持spc, spce, tip3p, tip4p, tip5p类型，默认是tip3p。
  - 离子种类：支持NaCl、MgCl<sub>2</sub>、None几种类型。若平衡电荷设置为True，则不支持离子种类选择None。
  - 离子浓度：离子浓度支持范围是0<离子浓度≤5mol/L，默认是0.15mol/L，如果离子种类选择了None，则不支持选择离子浓度。
3. 配置配体文件和相关参数信息。
  - 配体文件：仅支持mol2、pdb、sdf格式，且只支持3D结构，文件大小小于1M。若文件中含有多个配体，默认只处理第一个。
  - 配体力场：支持gaff/gaff2，蛋白力场gromacs力场与配体力场gaff/gaff2不兼容，如果蛋白力场选择了gromacs力场，则不支持上传配体。

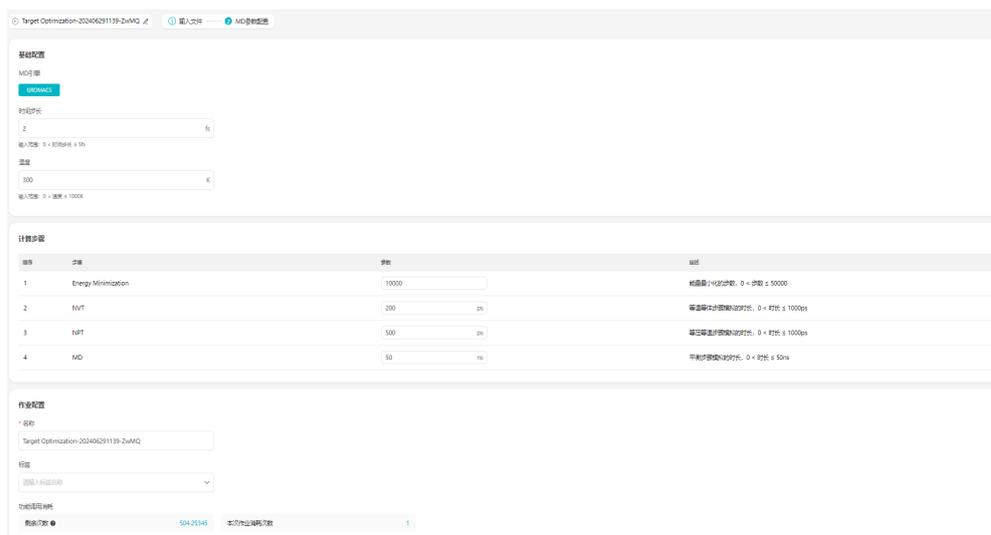
图 1-40 设置靶点优化配置页面



4. 单击“下一步”，配置靶点优化配置参数。
  - MD引擎：仅支持GROMACS。
  - 时间步长：MD模拟的时间步长，范围支持0<时间步长≤5fs，默认是2fs。
  - 温度：MD模拟的温度，范围支持0<温度≤1000K，默认是300K。
5. 配置MD的计算步骤
  - Energy Minimization：能量最小化的步数，范围支持0<步数≤50000，默认是10000步。
  - NVT：等温等体步骤模拟的时长，范围支持0<时长≤1000ps，默认是200ps。
  - NPT：等压等温步骤模拟的时长，范围支持0<时长≤1000ps，默认是500ps。

- MD: 平衡步骤模拟的时长, 范围支持0<时长≤50ns, 默认是50ns。
6. 配置作业配置, 单击“提交”。
- 名称: 设置作业名称。
  - 标签: 设置作业标签。
  - 功能调用消耗: 运行一次功能会消耗一次。

图 1-41 设置靶点优化 MD 配置页面



7. 查看靶点优化结果

单击作业名称, 进入作业结果页面, 拟时结果根据构象RMSD聚类。

- 单击“查看轨迹”, 下载运动轨迹。
- 单击“下游分析”, 将聚类之后的不同构象进行分子对接。
- 单击每一个Model的 按钮, 查看构象。
- 单击每个Model的 按钮, 可以收藏构象, 收藏后可直接在收藏夹页查看。

图 1-42 靶点优化结果页面



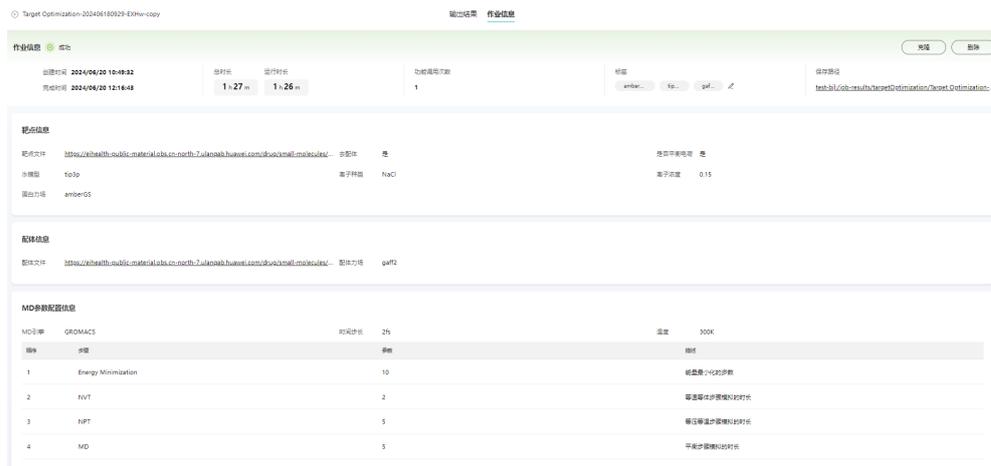
图 1-43 靶点优化结果查看轨迹



8. 查看靶点优化作业详情

在“作业中心”单击作业名称，单击“作业信息”页面。可以查看作业的参数和运行信息。

图 1-44 靶点优化结果详情



## 1.7.2 苗头化合物发现

### 1.7.2.1 分子对接

分子对接基于华为云大算力，可根据靶点蛋白和小分子药物的3D结构，计算对接结合能，实现百万级别虚拟筛选。

- 单击“分子对接”功能卡片，进入分子对接受体预处理页面，单击  上传受体文件，进行受体蛋白预处理配置。

受体文件仅支持PDB格式，若文件中存在多个受体，默认只处理第一个。受体文件支持上传1个或者多个文件，最多支持输入20个受体文件，文件来源包含数据中心、PDB库、示例数据。

图 1-45 上传受体文件



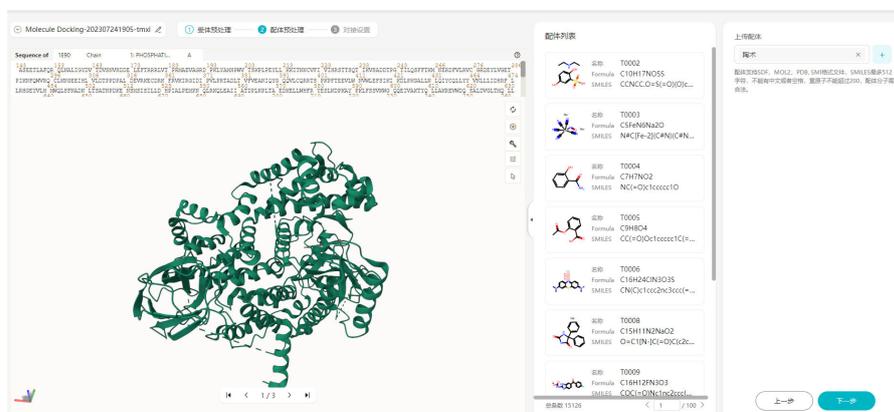
2. 受体预处理参数配置。
  - 去水：去掉受体蛋白里面的水分子。
  - 去金属离子：去掉受体蛋白里面的金属离子。
3. 单击“应用全局”，可以对受体进行批量处理。
4. 单击“下一步”，进入配体预处理页面。
5. 单击“上传配体”后的 +，选择配体文件。

配体文件仅支持输入1个文件。

配体支持SDF、MOL2、PDB、SMI格式文件，SMILES最多512字符，不能有中文或者空格，重原子不能超过200，配体分子需合法。

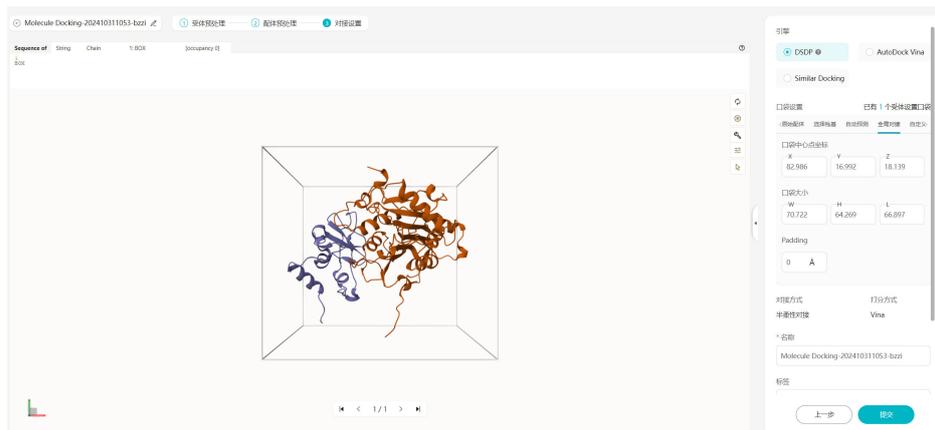
配体列表可显示总条数，但最多展示1000个小分子，对接时使用全量小分子。配体来源包含：数据中心、示例数据、公共库。且公共库预置了分子数据库，包含了陶术数据库、DrugSpaceX数据库、DrugBank数据库可以进行使用，单击“公共库”，可以进行选择。

图 1-46 选择配体文件



6. 单击“下一步”，在对接设置页面配置相关参数。

图 1-47 设置参数信息



- 引擎：对接引擎支持DSDP，AutoDock Vina和Similar Docking。DSDP是由北大高毅勤团队研发的基于GPU加速的对接方法，支持全局对接。  
如果对接引擎选择Similar Docking，不需设置口袋信息，但需要为每一个靶点上传一个“模板配体文件”。
  - 配置受体口袋位置，如果包含多个受体文件，需要每个受体都进行口袋设置，才能单击提交。
    - 原始配体  
将原始配体作为口袋位置，也可以通过上传配体文件来进行口袋设置。
    - 选择残基  
选择某些残基来作为口袋位置。
    - 自动预测  
如果受体口袋位置未知，可以使用自动预测软件来进行口袋位置预测。
    - 全局对接  
将整个受体作为口袋位置。
    - 自定义  
手动修改口袋位置和大小。
    - padding  
口袋位置放大多少尺寸。
7. 配置成功后，单击“提交”。
  8. 查看分子对接结果。  
如果是多受体对多配体，打开作业结果页面可以看到结合能二维矩阵，支持分别按照靶点和小分子进行排序。

图 1-48 查看输出结果 (1)



单击受体结合能框，跳转到单受体对多配体的结果表页面，可以下载全量及单条对接结果。

如果需要下载多个结果，可以选择结果后，单击左上角的“下载”。如果需要下载单条数据，单击数据操作列的“下载”即可。

图 1-49 查看输出结果 (2)

序号	分子图	名称	Formula	SMILES	Vina Score	Molecular Weight (MW)	logP	QED	Elastic	操作
621		Clidazemine	C27H22ClN4	CC(C)N=C1c2cnc...	-5.286	473.41	7.49	0.275	7.948	查看详情 查看3D 下载分析 下载3D
495		Ergotamine	C33H43N5O5	CN1CC#HBC1C...	-5.282	581.67	1.991	0.434	8.821	查看详情 查看3D 下载分析 下载3D
866		Dutasteride	C27H30F6N2O2	CC(C)12C=CC1=...	-5.038	528.54	6.576	0.427	8.548	查看详情 查看3D 下载分析 下载3D
981		Stigflatin	C18H15FN5O	NC(C)@H1CC1=O...	-4.944	407.32	2.016	0.622	7.108	查看详情 查看3D 下载分析 下载3D
650		Chlorhexidine	C22H30Cl2N10	N=C1NCCCCC1N...	-4.855	505.46	4.175	0.139	7.467	查看详情 查看3D 下载分析 下载3D

单击 可以按照属性以及相互作用力进行高级筛选。

图 1-50 高级筛选



单击“查看3D”可以看到小分子和靶点的对接构象。

图 1-51 查看 3D



单击配体对比，可以计算对接后的构象与输入小分子的RMSD值和输入小分子和靶点的相互作用2D图。

图 1-52 设置配体对

配体对比

原始配体

请选择

上传配体文件

请选择文件 +

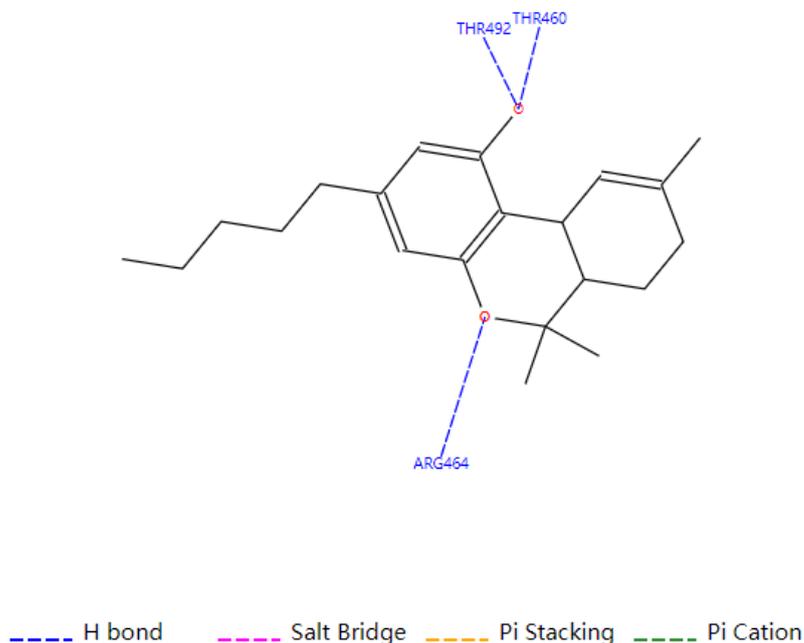
计算RMSD

图 1-53 配体对接结果图



单击查看2D相互作用图，可以看到对接后的小分子构象与蛋白的2D相互作用图，并且支持相互作用2D图下载。

图 1-54 查看 2D 相互作用图



单击“下游分析”选择分子搜索、分子属性预测、分子优化或分子属性预测或合成路径规划或自由能微扰后，单击“确定”即可创建下游分析。

单击“下载3D”可以下载小分子或复合物，小分子支持SDF和PDB格式，复合物只支持PDB格式。

单击“查看详情”可以直观查看小分子的结合能和属性信息。

图 1-55 查看详情

名称 Clofazimine  
Formula C<sub>27</sub>H<sub>22</sub>ClN<sub>4</sub>  
SMILES CC1CN=C1c2c2n(c3ccc(Cl)cc3)c3ccc(cc3n2)c1n1cccc(Cl)c1

**属性信息**

Medicinal Chemistry	Physicochemical Property
QED 0.275	Molecular Weight (MW) 473.41
SAscore 7.948	Molecular Refractivity (MR) 138.1357
SCScore 4.751	Check Acid base
Fsp <sup>3</sup> 0.11111	nHA 4
NPscore -0.336	nHD 1
Lipinski Rule false	nRot 6
Pfizer Rule false	nRing 5
GSK Rule false	MaxRing 6
Golden Triangle true	nHet 6
RAING 0 alerts(s)	nHDonor 0

Binding Affinity  
Vina Score 1 -5.286  
View Score 2 ...

单击 按钮，可以收藏对接结果，收藏后可直接在收藏夹页查看。

9. 分子对接结果支持以卡片视图的形式进行查看，参考图1-56，在卡片视图中：

图 1-56 查看输出结果 ( 3 )

分子ID	分子名称	Vina Score 1	QED	SaScore	Logo
3801	Belinostat	-6.102	0.446	7.411	2.000
3918	IRX-5183	-5.901	0.581	7.644	5.623
8509	Avacopan	-5.889	0.29	8.404	8.048
7141	HSD-016	-5.874	0.61	6.137	4.514
495	Ergotamine	-5.839	0.434	8.821	1.951
3417	Aminoquinuride	-5.637	0.42	7.502	4.208
3297	Genz-10850	-5.546	0.536	5.364	4.696
4029	Tolvaptan	-5.54	0.541	7.361	5.683
6998	Tekagepant	-5.532	0.471	7.877	3.686
4134	Enzastaurin	-5.51	0.337	7.289	4.932

- 单击右上方的选择下拉框，可以选择分子的排序方式，将分子按照所选的排序方式进行展示。

图 1-57 排序方式



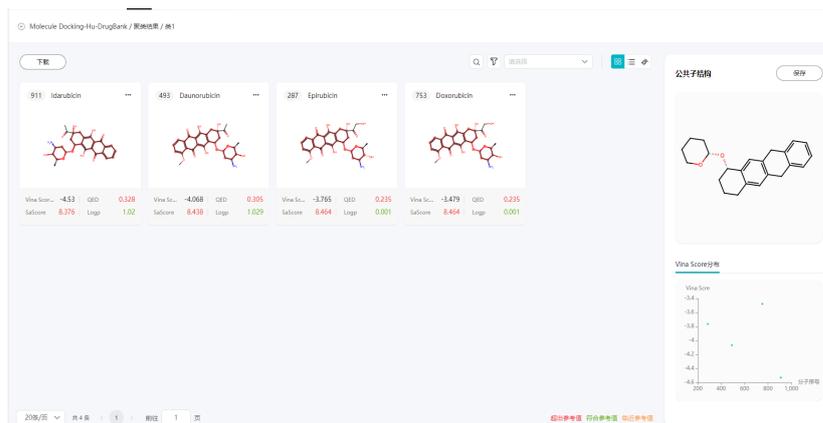
- 单击每个分子卡片右上方的“...”，可以选择“查看详情”、“查看3D”、“下游分析”、“下载3D”。
  - 查看详情：单击查看详情，跳转至分子列表页进行查看。
  - 查看3D：查看分子的3D视图。
  - 下游分析：分子优化对应的下游分析为分子搜索、分子属性预测、分子优化、自由能微扰和合成路径，单击“确定”即可创建。
  - 下载3D：如果分子设置了靶点，可以单击“下载3D”下载优化后的小分子或者复合物，如果分子未设置靶点，单击“下载3D”只能下载小分子，小分子支持SDF和PDB格式，复合物只支持PDB格式。
- 每个分子卡片上会展示相应分子序号与对应的参数Vina Score、QED、SaScore
  - Vina Score：代表分子如果添加了靶点，将会计算对接结合能，并按照Vina Score进行排序
  - QED：代表分子的成药性。
  - SaScore：代表合成可及性分数，旨在评估分子的合成难易程度。

#### 10. 查看作业信息

点击作业的“作业信息”页面可以查看作业的运行时间和配置参数。



图 1-60 查看聚类详情



### 说明

- 每个结果页面只用进行一次聚类分析操作。
- 聚类结果是存成文件，如果文件被删或者获取不到的话会有警告，聚类结果不存在。此时可以单击“重新聚类分析”。
- 如果聚类失败，根据提示失败原因解决问题后，可单击“重新聚类分析”。

## 相关参数

分子对接的具体参数，请参考[相关参数](#)。

### 1.7.2.2 分子属性预测

基于盘古药物分子大模型，预测化合物ADMET相关的80多种成药属性，有些属性的预测值会给出置信区间，更好地辅助分子设计。

1. 单击“分子属性预测”功能卡片，进入配置页面。

图 1-61 小分子配置页面

2. 在配置页面输入分子信息，及配置相关参数。
  - 输入方式：支持绘制分子、选择文件、手动输入。
    - 绘制分子：只能绘制一个分子，能够输入分子的SMILES。
    - 选择文件：选择分子文件，最多支持100万个小分子，且分子文件大小不超过2GB。支持SDF、MOL2、PDB、SMI格式。文件来源包括数据中心和示例数据。
    - 手动输入：输入小分子SMILES表达式。最多支持输入1000行，每行最多输入512个字符，SMILES不支持输入空格或者中文。
  - 选择基模型：选择基模型，此参数只有专业版支持。基模型列表见[AI建模](#)。
  - 选择属性模型：选择AI模型。如果需要创建模型，可参考[AI模型](#)。此参数只有专业版支持。一次最多可以选10个模型属性。属性模型的基模型必须与上一步所选择的基模型一致。
  - 输出小分子表征：是否输出小分子表征文件，小分子表征文件可以用作训练模型使用，默认是不输出。
  - 作业名称：设置作业名称。长度为5~64个字符，仅可以使用字母、数字、下划线“\_”、中划线“-”和空格，首位只能以数字或字母开头。
  - 标签：设置任务标签。
  - 功能调用消耗：20000个小分子记一次功能调用。

### 📖 说明

引用外部桶时，需要确保所引用的数据不超过45层级的目录。

3. 配置完成后，单击“提交”。
- 提交成功后，可以在“作业中心”查看执行结果。
4. 查看运行结果。

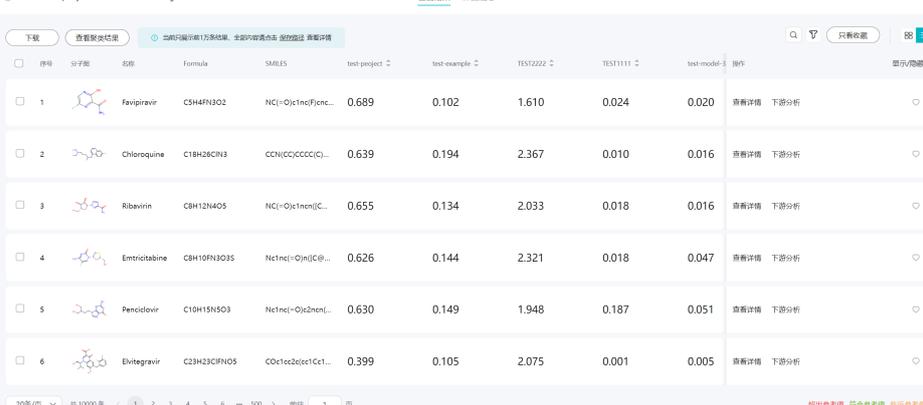
可以单击“下载”，下载分子属性预测结果信息。单击“操作”列的“查看”，查看分子信息。分子属性预测对应的下游分析为分子优化和合成路径，通过单击“下游分析”可以进行创建。

单击 ，可以收藏该分子属性预测结果，收藏后的结果可在收藏夹页直接查看。

### 📖 说明

下载操作将会产生流量费用，具体可参考[计费说明](#)。

图 1-62 查看结果 (1)

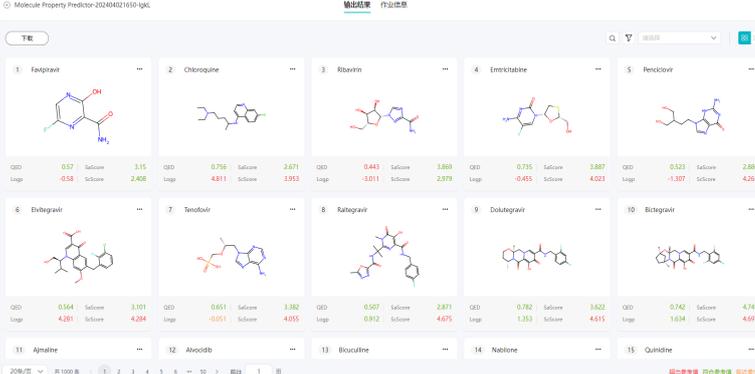


序号	分子图	名称	Formula	SMILES	test-project	test-example	TEST2222	TEST1111	test-model	操作
1		Favipiravir	C <sub>14</sub> H <sub>18</sub> N <sub>4</sub> O <sub>2</sub>	NC(=O)c1nc[nH]c1	0.689	0.102	1.610	0.024	0.020	查看详情 下游分析
2		Chloroquine	C <sub>18</sub> H <sub>26</sub> ClN <sub>3</sub>	CCN(C)CCCC(C)C	0.639	0.194	2.367	0.010	0.016	查看详情 下游分析
3		Ribavirin	C <sub>8</sub> H <sub>12</sub> N <sub>4</sub> O <sub>5</sub>	NC(=O)c1nc[nH]c1	0.655	0.134	2.033	0.018	0.016	查看详情 下游分析
4		Emtricitabine	C <sub>8</sub> H <sub>10</sub> F <sub>3</sub> N <sub>3</sub> O <sub>3</sub>	NC1=CN(C)C=C1	0.626	0.144	2.321	0.018	0.047	查看详情 下游分析
5		Penciclovir	C <sub>10</sub> H <sub>15</sub> N <sub>3</sub> O <sub>3</sub>	NC1=CN(C)C=C1	0.630	0.149	1.948	0.187	0.051	查看详情 下游分析
6		Eltigravir	C <sub>23</sub> H <sub>23</sub> ClF <sub>4</sub> N <sub>3</sub> O <sub>5</sub>	COc1cc2c(c1)C=C2	0.399	0.105	2.075	0.001	0.005	查看详情 下游分析

5. 分子属性预测结果支持以卡片视图的形式进行查看，参考图1-63，在卡片视图中：

- 单击右上方的选择下拉框，可以选择分子的排序方式，将分子按照所选的排序方式进行展示。

图 1-63 查看运行结果 (2)



名称	分子图	QED	LogP	SaScore	其他指标
1: Favipiravir		0.57	-0.58	3.15	2.408
2: Chloroquine		0.755	4.811	2.671	3.953
3: Ribavirin		0.445	-3.011	3.869	2.979
4: Emtricitabine		0.735	-0.455	3.887	4.023
5: Penciclovir		0.523	-1.307	2.885	4.261
6: Eltigravir		0.564	4.281	3.101	4.284
7: Tenofovir		0.651	-0.951	3.382	4.055
8: Raltegravir		0.507	0.912	2.871	4.675
9: Dolutegravir		0.782	1.353	3.622	4.815
10: Bictegravir		0.742	1.634	4.747	4.697
11: Almaline					
12: Almodib					
13: Bicucilline					
14: Nabillone					
15: Quinidine					

图 1-64 排序方式



- 单击每个分子卡片右上方的 ，可以收藏该分子属性预测结果，收藏后的结果可在收藏夹页直接查看。

- 单击每个分子卡片右上方的 “...” ，可以选择 “查看详情” 、 “下游分析” ” 。
  - 查看详情：单击查看详情，跳转至分子列表页进行查看。
  - 下游分析：分子优化对应的下游分析为分子优化、分子搜索和合成路径规划，单击 “确定” 即可创建。
- 每个分子卡片上会展示相应分子序号与对应的参数QED、SaScore
  - QED：代表分子的成药性。
  - SaScore：代表合成可及性分数，旨在评估分子的合成难易程度。

## 聚类分析

目前分子属性预测返回的结果小分子数较多，无法进行批量分析，通过一些聚类的辅助方式能更好的选择分子。从每个类里挑选出一两个分子进行后续分析和验证，提高分析的效率和分析质量。也可以通过聚类找出一些关键的骨架，来进行下游分析或者优化等。

1. 在输出结果页面左上角单击 “聚类分析” 后，系统开始进行分析，同时显示 “聚类分析中” 。

图 1-65 聚类分析



2. 待聚类分析完成后，单击 “查看聚类结果” 。进入聚类结果页。

图 1-66 查看聚类结果

聚类方法	分子数量
HQ MC	103

类序号	公共子结构	公共子结构原子数	分子数量	SMILES	操作
2		15	3	C1=CC(C@H)2C(C@H)3CCCC3C2)C1	查看详情
3		13	6	C1=CC=C(C2=CC=CC=C2)C=C1	查看详情
4		13	3	C1=CC=C(C2CCCC2)C=C1	查看详情
5		12	3	C1=CC=C(C2=CC=CC=C2)C=C1	查看详情
6		12	5	C1=CC=C(C(C)C)C2CCCC2)C=C1	查看详情

3. 在聚类结果页面，可以查看每个聚类的分子数量等信息。
4. 单击某个聚类的操作列的 “查看详情” ，即可进入聚类详情页面，聚类详情页支持以卡片、列表以及3D的形式查看。默认展示卡片页面，用户可自行进行切换。

## 📖 说明

- 每个结果页面只用进行一次聚类分析操作。
- 聚类结果是存成文件，如果文件被删或者获取不到的话会有警告，聚类结果不存在。此时可以单击“重新聚类分析”。
- 如果聚类失败，根据提示失败原因解决问题后，可单击“重新聚类分析”。

## 相关参数

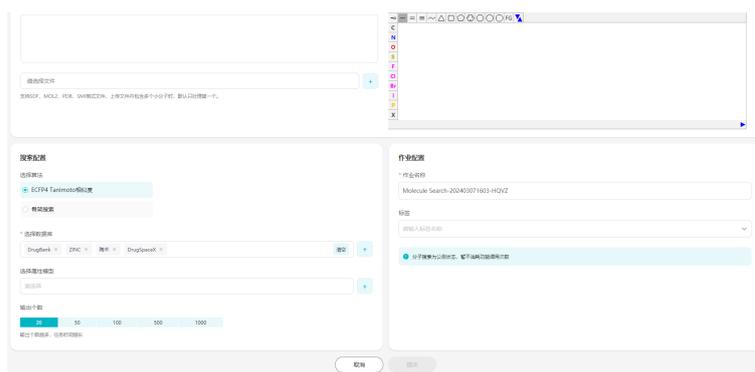
分子属性预测的具体参数请参考[相关参数](#)。

### 1.7.2.3 分子搜索

可自定义数据库，以用户输入的参考化合物结构为起点，可以按照相似度或者骨架从小分子数据库中搜索到相似结构和排序，可实现百亿级小分子的秒级搜索。

1. 单击“分子搜索”功能卡片，进入分子搜索页面。
2. 在配置页面，进行分子搜索配置，包括输入小分子、选择搜索算法、选择输出个数。

图 1-67 分子搜索页面



- 输入小分子：可以通过输入SMILES、上传文件或者直接绘制输入小分子。最终以SMILES为准。
  - 选择算法：可以选择ECFP4 Tanimoto相似度或者骨架搜索。ECFP4 Tanimoto相似度是通过ECFP4指纹计算Tanimoto相似度来搜索相似度比较高的小分子。骨架搜索是通过设置分子骨架搜索具有相同骨架的分子。
  - 选择数据库：最多可选择10个数据库。
  - 选择属性模型：选择AI模型。如果需要创建模型，可参考[AI模型](#)。此参数只有专业版支持。一次最多可以选10个模型属性。
  - 输出个数：输出个数越多，任务时间越长。
  - 作业名称：设置作业名称。长度为5~64个字符，仅可以使用字母、数字、下划线“\_”、中划线“-”和空格，首位只能以数字或字母开头。
  - 标签：设置任务标签。
3. 单击“提交”。可在作业中心查看该作业的运行情况。运行完成后，可在作业中心单击该作业查看输出结果。

单击  按钮，可以收藏分子搜索结果，收藏后的结果可直接在收藏夹页查看。

图 1-68 作业中心

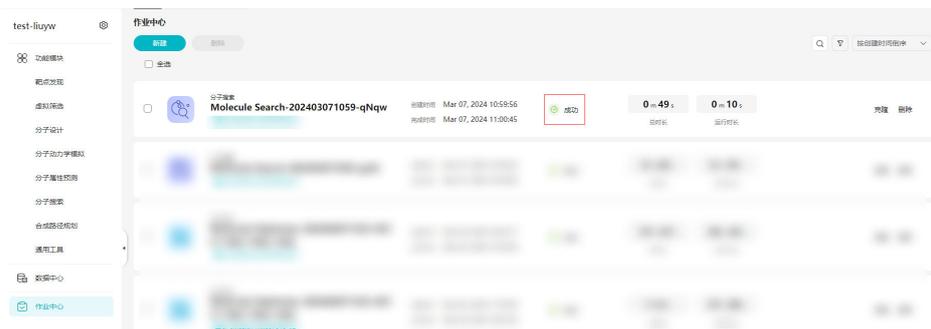


图 1-69 输出结果 (1)

Molecule Search-202408101451-8MST

输出结果 作业信息

ID	分子图	名称	相似度	Formula	SMILES	Similarity	Molecular Weight (kDa)	logP	QED	SA	操作
0		Original	0	C17H18ClN3O	Ck1ecczo(c1c)N...	1	313.79	3.429	0.809	2	查看详情 下游分析
1		DE00543	3	C17H18ClN3O	Ck1ecczo(c1c)N...	1	313.79	3.429	0.809	2	查看详情 下游分析
2		DE6209229	1	C18H19ClN3O	Ck1ecczo(c1c)N...	0.909	327.81	3.819	0.799	2	查看详情 下游分析
3		DE8072628	1	C18H19ClN4O	Ck1ecczo(c1c)N...	0.907	342.83	2.976	0.772	2	查看详情 下游分析
4		DE49035	1	C19H20ClN3O	Ck1ecczo(c1c)N...	0.889	341.84	4.209	0.78	2	查看详情 下游分析
5		DE8072627	1	C18H19ClN4O	Ck1ecczo(c1c)N...	0.867	342.83	2.976	0.772	2	查看详情 下游分析

图 1-70 输出结果 (2)

Molecule Search-202408101451-8MST

输出结果 作业信息

ID	分子图	名称	相似度	Database
0		Original	1	DrugBank
1		DE00543	1	DrugBank
2		DE6209229	0.909	DrugSpaceX
3		DE8072628	0.907	DrugSpaceX
4		DE49035	0.889	DrugSpaceX
5		DE8072627	0.867	DrugSpaceX
6		DE8072637	0.83	DrugSpaceX
7		DE8072635	0.83	DrugSpaceX
8		DE8072756	0.818	DrugSpaceX
9		DE8072636	0.813	DrugSpaceX

图 1-71 查看分子详情

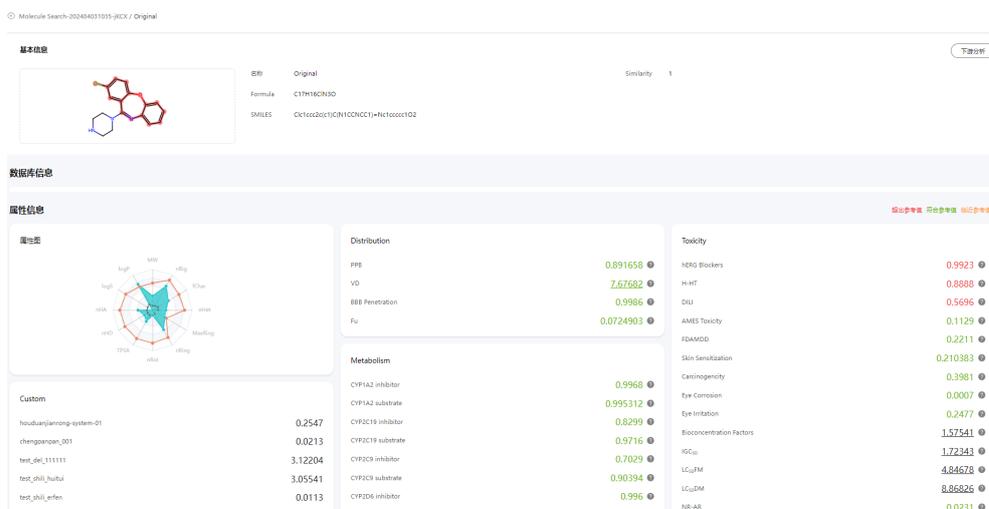


图 1-72 分子下游分析



图 1-73 作业信息页面



### 1.7.2.4 CPI 预测

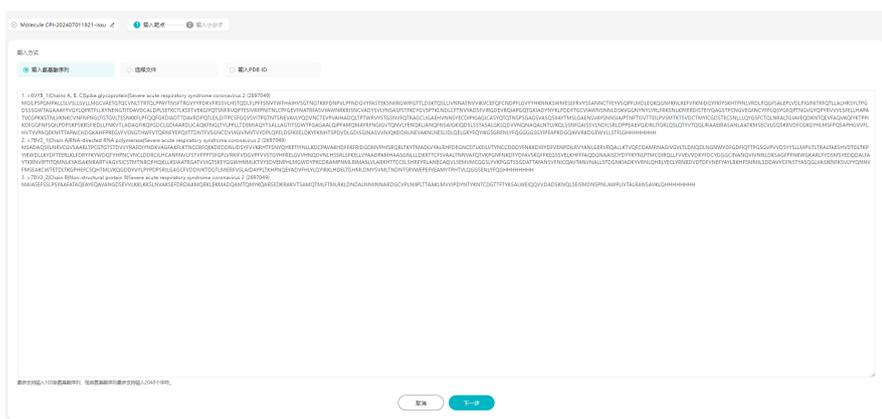
CPI预测基于蛋白质的一级序列和化合物的2D结构进行靶点匹配，精确的预测化合物-蛋白相互作用。

1. 单击“CPI预测”功能卡片，进入配置页面。
2. 配置靶点文件。

支持3种输入方式，分别是输入氨基酸序列、选择文件、输入PDB ID

- 输入FASTA格式氨基酸序列，输入框最多支持输入100条氨基酸序列，每条氨基酸序列最多支持输入2048个字符。

图 1-74 靶点配置输入氨基酸序列



- 选择文件，支持fasta格式和pdb文件。
  - fasta格式文件，最多上传1个文件，文件中最多支持100条氨基酸序列，每个靶点的氨基酸序列长度不超过2048。
  - pdb格式文件，最多上传100个靶点文件，如果上传多聚体靶点，仅解析第一条链，每个靶点的氨基酸序列长度不超过2048。

图 1-75 靶点配置选择文件



- 输入PDB ID，最多输入100个PDB ID，用逗号或换行方式隔开；不区分大小写；如果输入多聚体靶点，仅解析第一条链；每个靶点的氨基酸序列长度不超过2048。

图 1-76 靶点配置输入 PDB ID



3. 在配置页面输入分子信息，及配置相关参数。

- 输入方式：支持绘制分子、选择文件、手动输入。
    - 绘制分子：只能绘制一个分子，能够输入分子的SMILES。
    - 选择文件：选择分子文件，最多支持100万个小分子，且分子文件大小不超过2GB。支持SDF、MOL2、PDB、SMI格式。文件来源包括数据中心和示例数据。
    - 手动输入：输入小分子SMILES表达式。最多支持输入1000行，每行最多输入1024个字符，SMILES不支持输入空格或者中文。
  - 作业名称：设置作业名称。长度为5~64个字符，仅可以使用字母、数字、下划线“\_”、中划线“-”和空格，首位只能以数字或字母开头。
  - 标签：设置任务标签。
  - 功能调用消耗：100000个小分子记一次功能调用。
4. 配置完成后，单击“提交”。
- 提交成功后，可以在“作业中心”查看执行结果。
5. 查看多对多运行结果。
- 如果是多受体对多配体，打开作业结果页面可以看到结合能二维矩阵，支持分别按照靶点和配体进行排序。

图 1-77 查看结果 ( 1 )



6. 查看一对多运行结果。
- 单击受体结合能框，跳转到单受体对多配体的结果表页面，可以下载全量及单条CPI预测结果。
- 如果需要下载多个结果，可以选择结果后，单击左上角的“下载”。如果需要下载单条数据，单击数据操作列的“下载”即可。

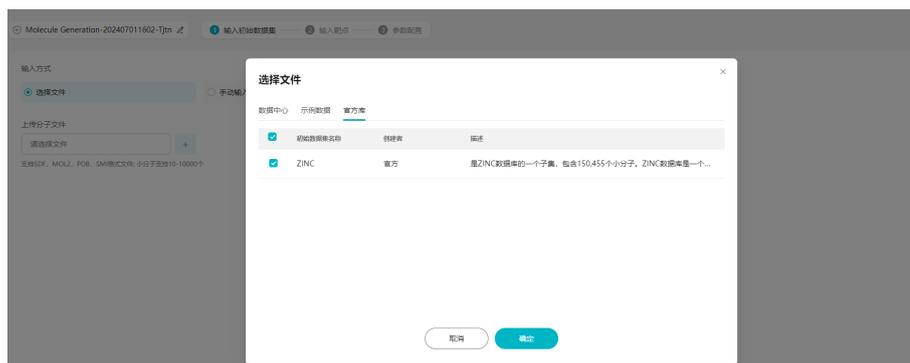
### 📖 说明

下载操作将会产生流量费用，具体可参考[计费说明](#)。

单个受体对多个配体的结果页面有列表视图、卡片视图，支持搜索、高级筛选，排序等功能。每个小分子支持“查看详情”可以进入小分子的属性详情页；支持“下游分析”，可以进行分子属性预测、分子搜索、分子优化、合成路径规划分析。支持收藏功能，单击  按钮，即可收藏CPI预测结果，收藏后可直接在收藏夹查看。



图 1-81 输入初始数据集



- 手动输入：最少输入10行，最多输入10000行，每行字符不超过512；SMILES不支持输入空格或者中文
3. 单击“下一步”，进入靶点设置，此步骤为可选步骤，如果需要设置靶点，并且将对接结合能作为一个约束条件进行分子生成，需要进行配置。最多可添加2个靶点。
- 如果不需要设置靶点，此步骤可以进行省略，如果设置了靶点，作业运行时间会加长。
- a. 通过“靶点设置”上传靶点，并且设置对接口袋。
 

此处靶点设置为可选参数，如果选择靶点设置，可以将对接活性作为一个约束条件进行分子生成。靶点1对应的约束条件是target1\_binding\_energy，靶点2对应的约束条件是target2\_binding\_energy

您可以通过“上传靶点”，将受体文件上传。受体文件仅支持pdb格式的文件，提交任务时系统会自动删除受体中包含水、配体和金属离子。删除受体文件后，运行任务时会跳过靶点设置步骤。
  - b. 口袋设置。
    - 原始配体：
 

将原始配体作为口袋位置，也可以通过上传配体文件来进行口袋设置。
    - 选择残基
 

选择某些残基来作为口袋位置。
    - 自动预测
 

如果没有受体口袋位置未知，可以使用自动预测软件来进行口袋位置预测。
    - 全局对接
 

将整个受体作为口袋位置。
    - 自定义
 

手动修改口袋位置和大小。
    - padding
 

口袋位置放大多少尺寸。
  - c. 对接引擎类型：DSDP、AutoDock Vina

图 1-82 靶点设置

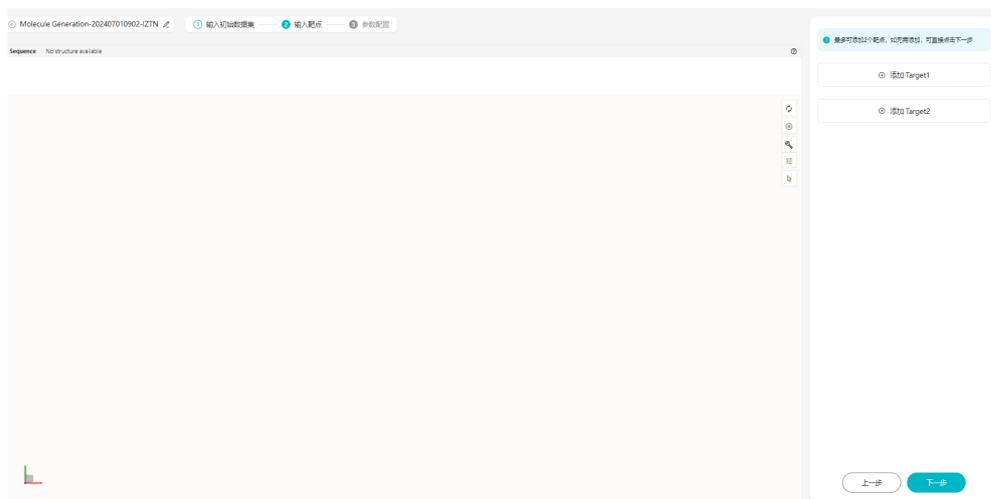


图 1-83 Target1 设置

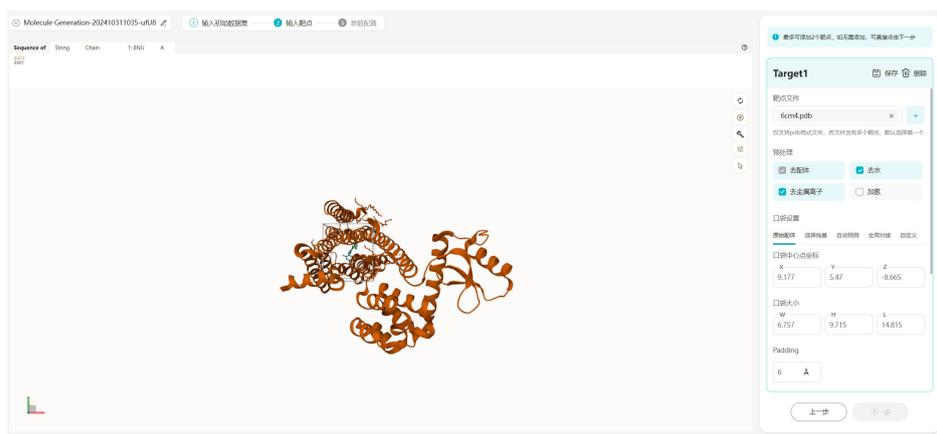


图 1-84 Target2 设置

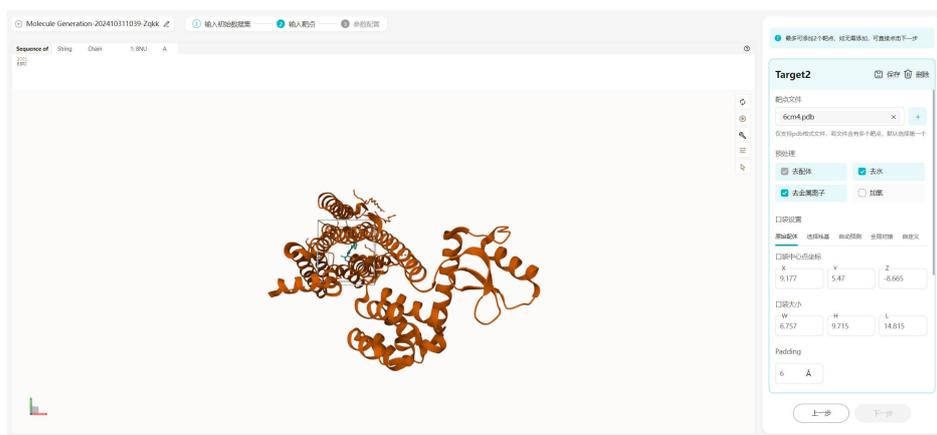
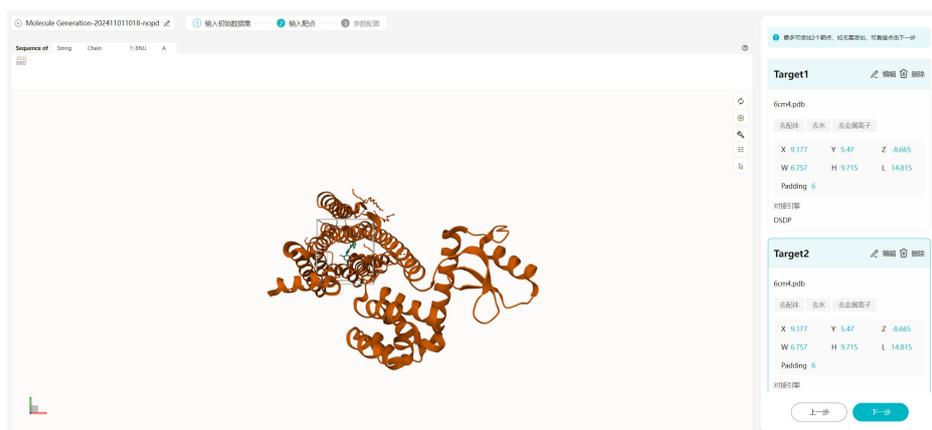


图 1-85 靶点设置完成



4. 单击“下一步”，进入参数设置页面。

- 选择基模型：支持选择基模型。此参数仅专业版支持。如果选择的基模型是非官方盘古药物大模型，则约束条件不支持官方机器学习属性，只支持以该基模型创建的属性模型作为约束条件。基模型列表见[AI建模](#)。
- 选择属性模型：选择AI模型。如果需要创建模型，可参考[AI模型](#)。此参数只有专业版支持。一次最多可以选10个模型属性。属性模型的基模型必须与上一步所选择的基模型一致。
- 设置强约束、弱约束的参数和参数值，相关参数含义见[相关参数](#)章节。
  - 强约束：生成的小分子必须要满足的约束条件。强约束用于严格筛选输出分子，若分子不满足其中任何一条约束则会被直接丢弃。强约束条件个数1~3个最佳，不宜过多，过多的强约束会导致输出结果数量少于预期或者没有分子生成。如果设置的官能团结构太大，建议放到弱约束，因为这样设置会使模型可探索的区间比较小，导致可能没有结果生成。
  - 弱约束：生成的小分子不必要满足的约束条件。弱约束用于打分排序最终结果，输出分子会按照弱约束顺序尽可能多地满足所有弱约束；不满足弱约束的结果依然会被保留在结果中，对于一个比较关注且重要的分子属性，建议放到强约束条件设置里，因为弱约束不会做过滤。

如果进行了“靶点设置”，会自动加上相应靶点的“Binding Free Energy”才会计算对接结合能，并将对接结合能作为约束条件进行分子生成。

您可以通过“添加”来添加新的约束条件，也可以在操作列单击删除图标，删除约束参数。每个约束参数的含义参考[相关参数](#)。

强约束最多选择5个，仅Substructure和Interaction参数可重复选择，其余不可以；弱约束最多选择10个，可以重复选择，弱约束的初始权重与弱约束的顺序有关，弱约束条件越靠前则初始权重越大。

- 输出个数：选择输出个数，目前支持500、1000、5000。输出个数越多，任务时间越长。
- 名称：可修改，修改后左上角也同步修改。长度为5~64个字符；仅可以使用字母、数字、下划线“\_”、中划线“-”和空格；首位只能以数字或字母开头。
- 标签：设置任务标签。
- 功能调用消耗：运行一次功能会消耗一次。

## 说明

生成后的小分子在满足强约束条件的基础上，会根据满足弱约束条件的权重总和以及参考小分子的相似度来打分并进行排序。在初始化权重的基础上，每个约束所占的权重，会在每一轮的分子生成迭代中，根据所满足的约束来进行动态调整。比如说约束条件1，在分子生成迭代中比较容易满足，那么该条件的权重会降低，如果不容易满足，该条件的权重会升高。

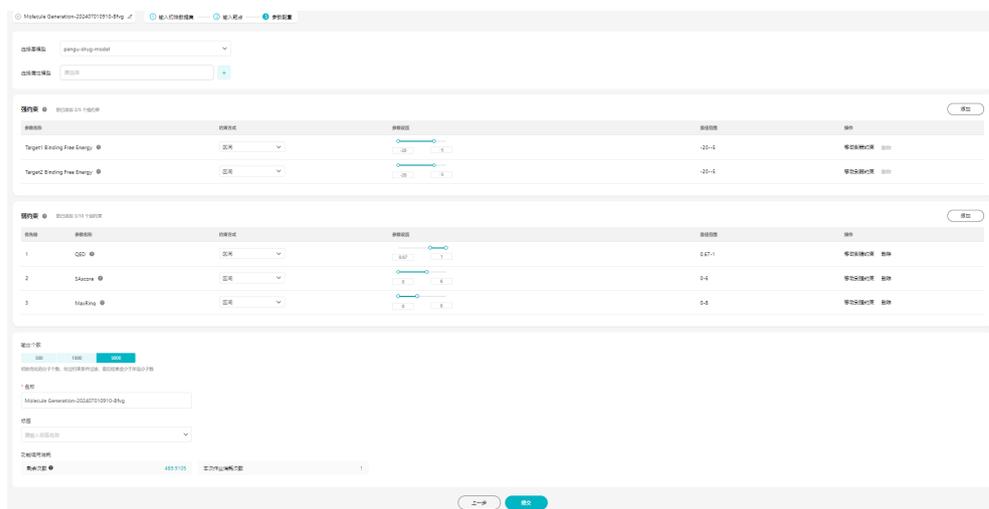
如果需要设置官能团的约束，可以在约束条件中设置“Substructure”，然后选择“包含”或者“排除”，包含官能团或者去除官能团条件，设置1个官能团数为佳。在官能团设置中，可以单击蓝色按钮来设置生长方向，即我们所生成的分子只会往我们所标记的方向生长。如果在一个Substructure约束条件里面添加多个官能团，则官能团之间的关系是“或”，即生成的分子满足其中一个官能团即可。如果添加多个Substructure约束，则官能团之间的关系是“和”，即生成的分子都会满足这几个官能团。

如果需要设置相互作用力的约束，可以在约束条件中设置“Interaction”，然后选择“包含”或者“排除”，包含相互作用力或者去除相互作用力。在相互作用力约束条件设置中，可以选择相应的靶点、氨基酸和相互作用力，相互作用力支持H bond, Hydrophobic, Salt Bridge, Pi Stacking, Pi Cation。如果在一个Interaction约束条件里面添加多个相互作用力，则相互作用力之间的关系是“或”，即生成的分子满足其中一个相互作用力即可。如果添加多个Interaction约束，则相互作用力之间的关系是“和”，即生成的分子都会满足这几个相互作用力。

约束方式，包含“区间”，“最大化”和“最小化”种方式，“区间”指的是我们所选择的属性包含在我们所设置的参数设置区间内。“最大化”指的是我们所设置的属性越大越好，而不只是限定在某区间内，当我们无法判断属性应该设置在什么区间，但是属性越高，成药性越好，可以设置属性设置条件为最大化，这个只能在弱约束里面进行设置。“最小化”指的是所设置的属性越小越好，与最大化相反，也是只能在弱约束里面进行设置。

相似度分数，是利用ECFP4分子指纹计算生成后分子与原始分子的Tanimoto相似性。我们设置了禁止优化列表，禁止以高毒性为优化目标的属性优化，列表为：hERG Blockers, H-HT, DILI, AMES, Skin Sensitization, Carcinogenicity, Eye Irritation, Eye Corrosion。

图 1-86 参数设置页面



5. 单击“提交”，提交任务。
6. 分子生成结果支持以列表视图的形式进行查看
  - 可以以列表的形式查看分子生成的作业，单击左上角“下载”，下载分子生成的结果或者分子3D构象。如果分子设置了靶点，可以下载小分子或复合物，若分子未设置靶点，只能下载小分子。小分子支持SDF和PDB格式，复合物只支持PDB格式。

- 分子生成对应的下游分析为分子搜索、分子优化和合成路径规划，如果分子设置了靶点，可以选择自由能微扰进行下游分析，通过单击“下游分析”可以进行创建。
- 如果添加了靶点，支持按照相互作用力进行高级筛选，单击  进行条件配置。
- 单击  可以收藏分子生成结果，收藏的结果可在收藏夹页直接查看。

图 1-87 查看结果（1）

序号	分子图	名称	Formula	SMILES	Score	chengpanpan_001	Molecular Weight (Da)	logP	QED	操作
1		mol_1_C16...	C16H13O2N3O3S	Cc1ccc(C1=O)N2C...	0	0.183	398.27	3.528	0.586	查看详情 查看3D 下游分析 下载3D
2		mol_2_C21...	C21H19F2N3O4	CC(C)C(=O)C1=O)N...	0	0.006	415.4	3.86	0.808	查看详情 查看3D 下游分析 下载3D
3		mol_3_C18...	C18H20F3N3O5	CC(C)C1=O)N(C1=O)...	0	0.086	387.35	2.98	0.555	查看详情 查看3D 下游分析 下载3D
4		mol_4_C20...	C20H16N4O2S	CC1=O)N1C1=O)C...	0	0.011	376.44	3.499	0.681	查看详情 查看3D 下游分析 下载3D
5		mol_5_C15...	C15H16B1N3O4	COCCOC1=CC1=CC...	0	0.320	382.21	2.388	0.416	查看详情 查看3D 下游分析 下载3D
6		mol_6_C22...	C22H22N4O5	COc1cc(C)nc2ccc...	0	0.592	390.51	4.535	0.52	查看详情 查看3D 下游分析 下载3D

图 1-88 高级筛选

高级筛选 ×

筛选条件: 1.000 ≤ Score ≤ 3.000; Interaction: Target1-A:ASN:1093-H bond;

筛选维度 Score 1.000 - 3.000

低 高

Inter... ▼

靶点      氨基酸      相互作用力

Target1 ▼    A:ASN:1093 ▼    H bond ▼

⊕

添加筛选条件

### 说明

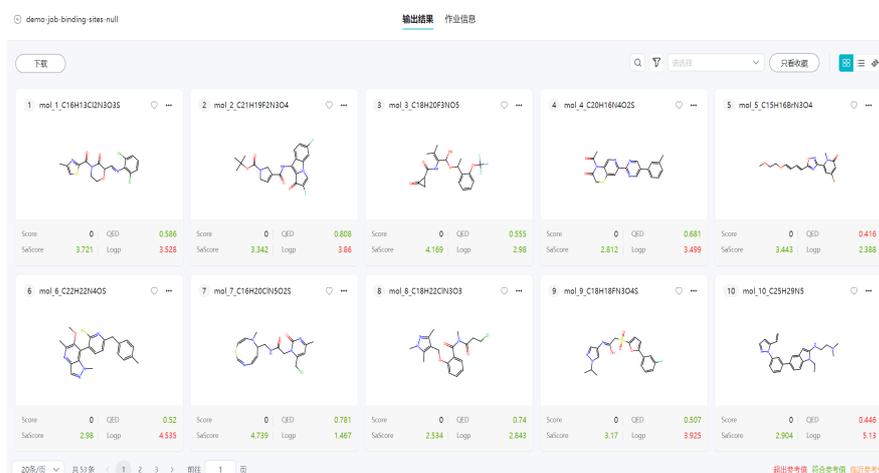
- 下载操作会产生流量费用，具体可参考[计费说明](#)。
- 分子生成结果支持以卡片视图的形式进行查看，参考图1-90，在卡片视图中：
    - 单击右上方的选择下拉框，可以选择分子的排序方式，将分子按照所选的排序方式进行展示。

图 1-89 排序方式



- 单击每个分子卡片右上方的  可以收藏分子生成结果，收藏的结果可在收藏夹页直接查看。
- 单击每个分子卡片右上方的  ，可以选择“查看详情”、“查看3D”、“下游分析”、“下载3D”。
  - 查看详情：单击查看详情，跳转至分子详情页进行查看。
  - 查看3D：查看分子的3D视图。
  - 下游分析：分子生成对应的下游分析为分子搜索、分子优化和合成路径，如果分子设置了靶点，可以选择自由能微扰进行下游分析，单击“确定”即可创建。
  - 下载3D：如果分子设置了靶点，可以单击“下载3D”下载生成的小分子或者复合物，如果分子未设置靶点，单击“下载3D”只能下载小分子。小分子下载支持SDF和PDB格式，复合物下载只支持PDB格式。
- 每个分子卡片上会展示相应分子序号与对应的参数Vina Score（有靶点）、Score、QED、SaScore
  - Vina Score：代表分子如果添加了靶点，将会计算对接结合能，并按照Vina Score进行排序
  - Score：代表生成小分子的综合打分
  - QED：代表分子的成药性。
  - SaScore：代表合成可及性分数，旨在评估分子的合成难易程度。

图 1-90 查看结果（2）



## 8. 分子生成结果支持以3D视图的形式进行查看

单击“查看3D”，可以看到分子的3D构象，如果设置了靶点，还可以看到生成的小分子与靶点的结合构象。

如果上传了双靶点，可以通过切换  $\langle 1/2 \rangle$  来切换靶点，查看相应靶点和生成分子的结合构象。如果设置了两个靶点会默认下载两个靶点的结果。

图 1-91 查看 3D 图

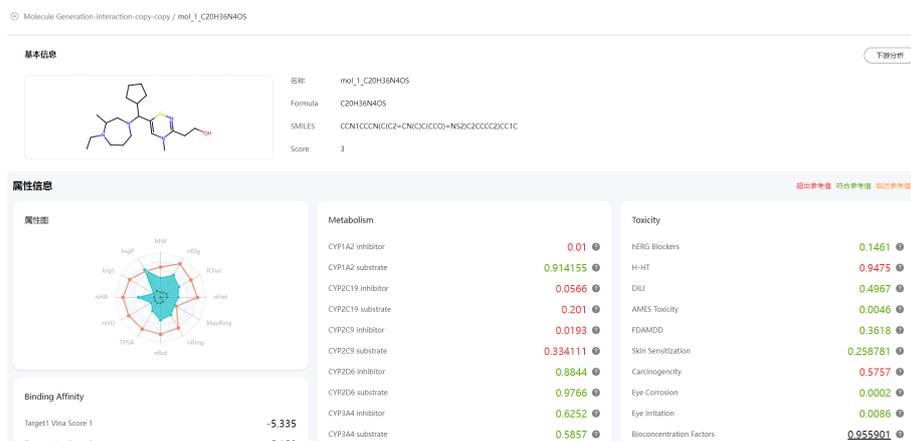


在查看3D的页面中，单击右侧的配体列表中，每个配体卡片右上角的“\*\*\*”，可以查看

- 查看详情：可以查看每个分子的属性信息和score。
- 查看属性：查看每个分子的基本属性。
- 查看2D相互作用图：查看靶点和分子之间的2D相互作用图，并且可以进行图片下载。
- 下游分析：分子生成对应的下游分析为分子搜索、分子优化和合成路径规划，如果分子设置了靶点，可以选择自由能微扰进行下游分析，单击“确定”即可创建。
- 下载3D：选择下载小分子或者复合物后，单击“确定”，小分子下载支持SDF和PDB格式，复合物只支持PDB格式。

## 9. 查看分子详情

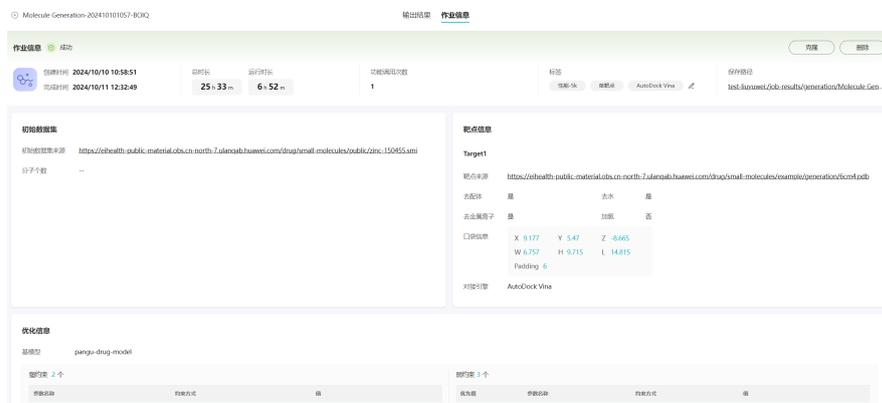
图 1-92 查看分子详情



## 10. 查看作业信息

单击“作业信息”切换到作业信息页签，可以查看作业的初始数据集、靶点信息与参数信息等。

图 1-93 作业信息



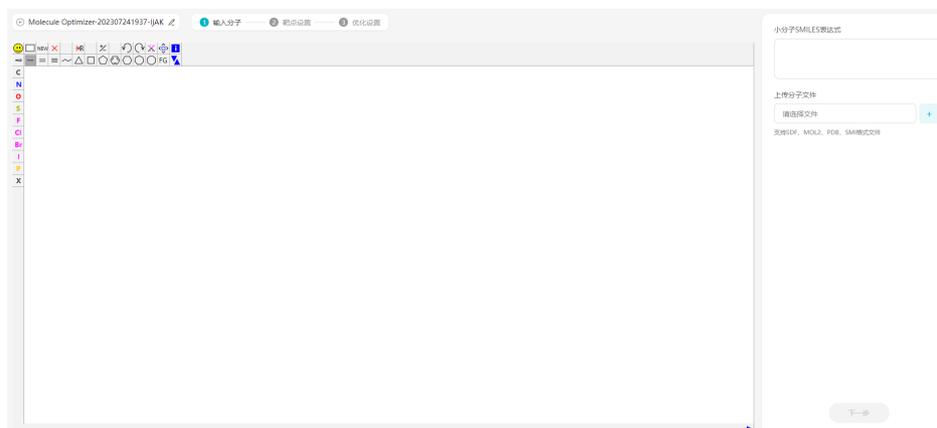
## 1.7.3 先导化合物优化

### 1.7.3.1 分子优化

分子优化基于盘古药物分子大模型，以参考化合物为起点，针对给定的期望理化性质，得到性质更优、结构新颖、与靶点蛋白亲和力高的化合物。

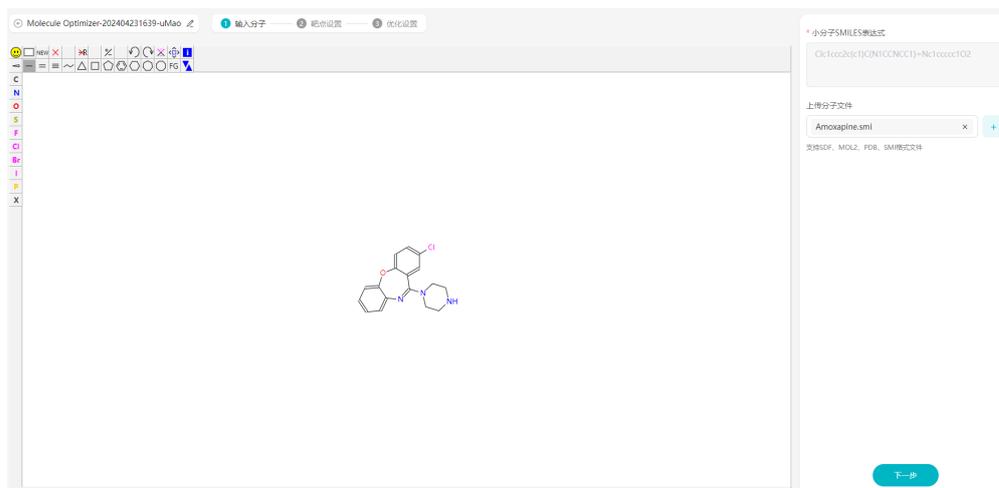
1. 单击“立即使用”，进入输入分子页面。

图 1-94 输入分子页面



2. 在白框内输入需要优化的小分子SMILES表达式或者上传分子文件。上传的文件支持SDF、MOL2、PDB、SMI格式。

图 1-95 输入分子



3. 单击“下一步”，进入靶点设置，此步骤为可选步骤，如果需要设置靶点，并且将对接结合能作为一个约束条件进行优化，需要进行配置。最多可添加2个靶点。在添加靶点后，会进行单分子预对接，可查看Top1的对接打分与构象。如果不需要设置靶点，此步骤可以进行省略，如果设置了靶点，作业运行时间会加长。

图 1-96 靶点设置

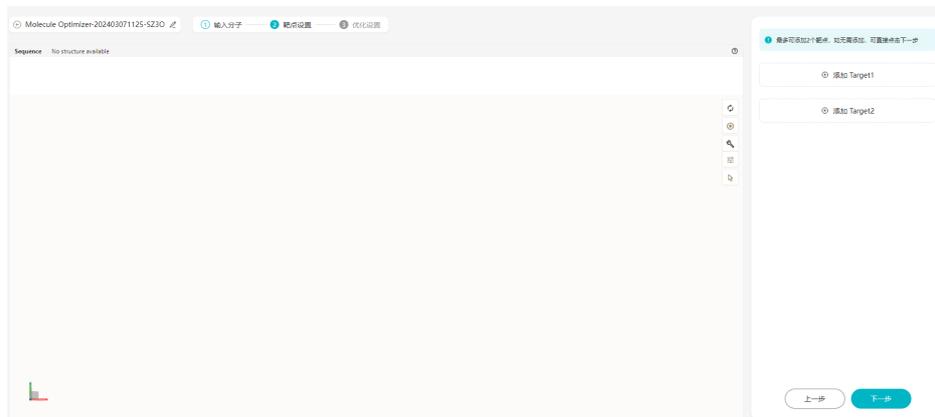


图 1-97 Target1 设置

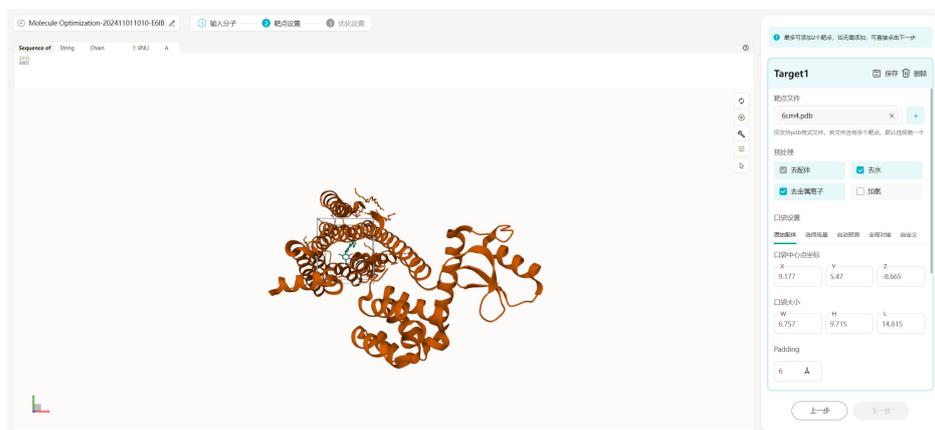


图 1-98 Target2 设置

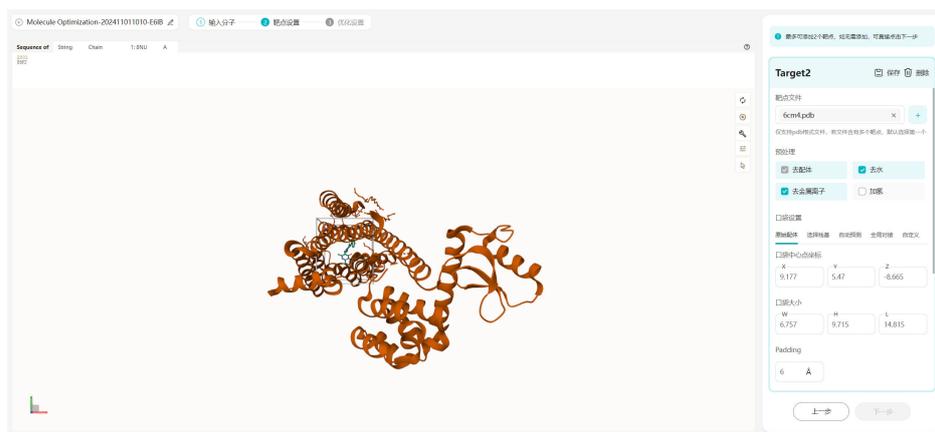


图 1-99 靶点设置完成



a. 通过“靶点设置”上传靶点，并且设置对接口袋。

此处靶点设置为可选参数，如果选择靶点设置，可以将对接活性作为一个约束条件进行分子优化。靶点1对应的约束条件是target1\_binding\_energy，靶点2对应的约束条件是target2\_binding\_energy

您可以通过“上传靶点”，将受体文件上传。受体文件仅支持pdb格式的文件，提交任务时系统会自动删除受体中包含水、配体和金属离子。删除受体文件后，运行任务时会跳过靶点设置步骤。

b. 口袋设置。

- 原始配体：

将原始配体作为口袋位置，也可以通过上传配体文件来进行口袋设置。

- 选择残基

选择某些残基来作为口袋位置。

- 自动预测

如果没有受体口袋位置未知，可以使用自动预测软件来进行口袋位置预测。

- 全局对接

将整个受体作为口袋位置。

- 自定义  
手动修改口袋位置和大小。
- padding  
口袋位置放大多少尺寸。

c. 对接引擎类型：DSDP、AutoDock Vina。

4. 单击“下一步”，进入优化设置页面。

图 1-100 优化设置页面（1）

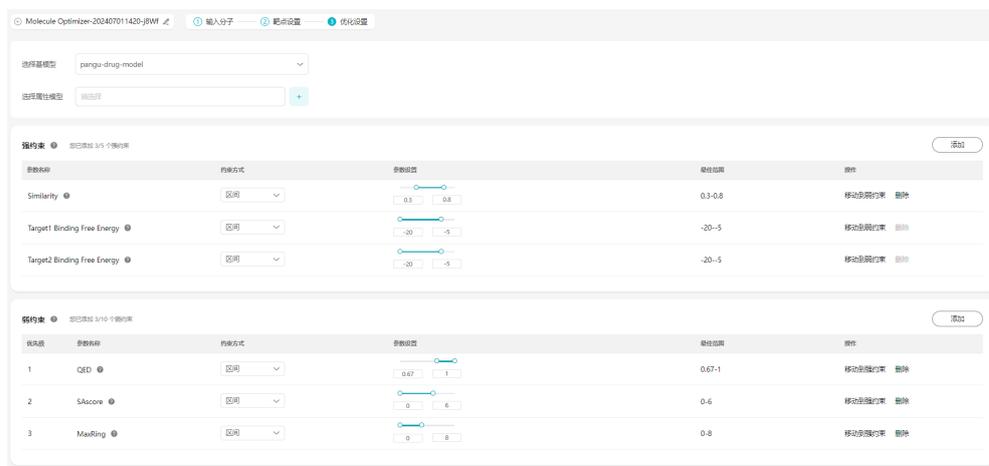
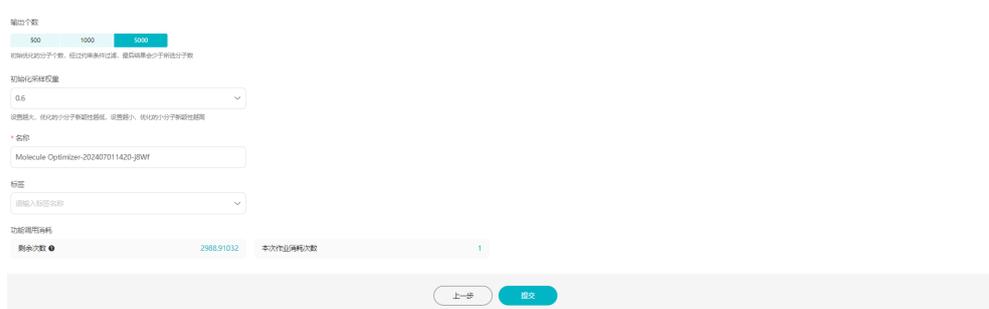


图 1-101 优化设置页面（2）



5. 单击“提交”。

6. 查看运行结果

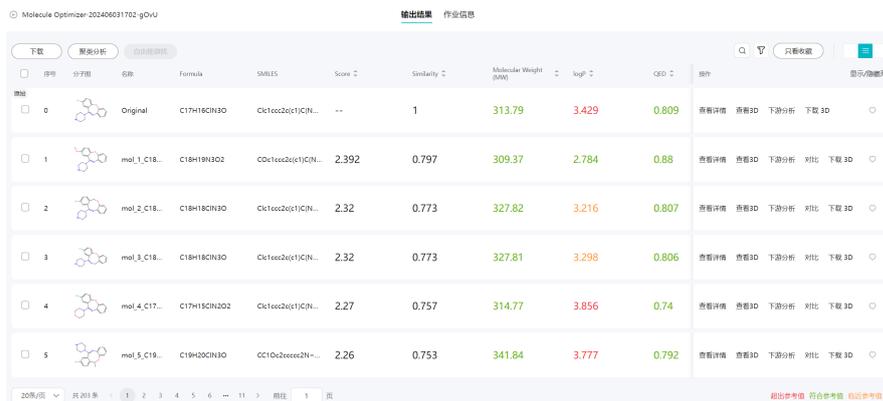
- 可以以列表的形式查看分子优化的作业，单击左上角“下载”，下载分子优化的结果或者分子3D构象。如果分子设置了靶点，可以下载小分子或复合物，若分子未设置靶点，只能下载小分子。小分子支持SDF和PDB格式，复合物只支持PDB格式。
- 单击  可以收藏结果，收藏后可直接在收藏夹页查看。
- 分子优化对应的下游分析为分子优化、分子搜索和合成路径规划，如果分子设置了靶点，可以选择自由能微扰进行下游分析，通过单击“下游分析”可以进行创建。

- 如果分子设置了靶点，可以单击“下载3D”下载优化后的小分子或者复合物，如果分子未设置靶点，单击“下载3D”只能下载小分子。小分子支持SDF和PDB格式，复合物只支持PDB格式。
- 如果添加了靶点，支持按照相互作用力进行高级筛选，单击进行条件配置。

### 说明

下载操作会产生流量费用，具体可参考[计费说明](#)。

图 1-102 查看结果 ( 1 )



ID	名称	Formula	SMILES	Score	Similarity	Molecular Weight (g/mol)	logP	CED	操作
0	Original	C17H18CN3O	Cc1ccc2c(c1)CN...	1	1	313.79	3.429	0.809	查看详情 查看3D 下游分析 下载3D
1	mol_1_C18...	C18H19N3O2	Cc1ccc2c(c1)CN...	2.392	0.797	309.37	2.784	0.88	查看详情 查看3D 下游分析 对比 下载3D
2	mol_2_C18...	C18H18CN3O	Cc1ccc2c(c1)CN...	2.32	0.773	327.82	3.216	0.807	查看详情 查看3D 下游分析 对比 下载3D
3	mol_3_C18...	C18H18CN3O	Cc1ccc2c(c1)CN...	2.32	0.773	327.81	3.298	0.806	查看详情 查看3D 下游分析 对比 下载3D
4	mol_4_C17...	C17H15CN2O2	Cc1ccc2c(c1)CN...	2.27	0.757	314.77	3.856	0.74	查看详情 查看3D 下游分析 对比 下载3D
5	mol_5_C19...	C19H20CN3O	CC1Oc2cccnc2N...	2.26	0.753	341.84	3.777	0.792	查看详情 查看3D 下游分析 对比 下载3D

图 1-103 高级筛选



高级筛选

筛选条件: 1.000 ≤ Score ≤ 3.000; Interaction: Target1-A:ASN:1093-H bond;

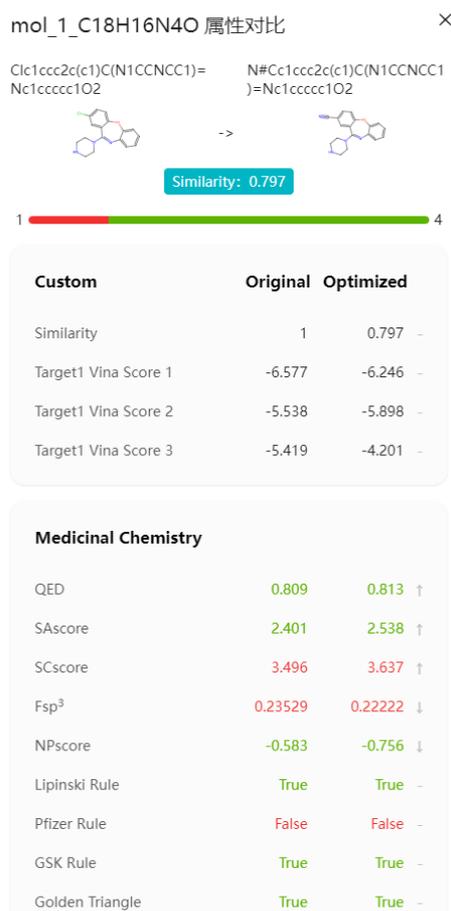
筛选维度: Score (1.000 - 3.000)

交互: Target1 (A:ASN:1093) H bond

添加筛选条件

7. 分子优化的结果可以查看原始配体和生成后的配体的结构以及属性值对比，可单击“对比”进行查看。可以直观地看到哪些属性变好了，哪些属性变坏了。

图 1-104 属性对比



8. 分子优化结果支持以卡片视图的形式进行查看，参考图1-106，在卡片视图中：
- 单击右上方的选择下拉框，可以选择分子的排序方式，将分子按照所选的排序方式进行展示。

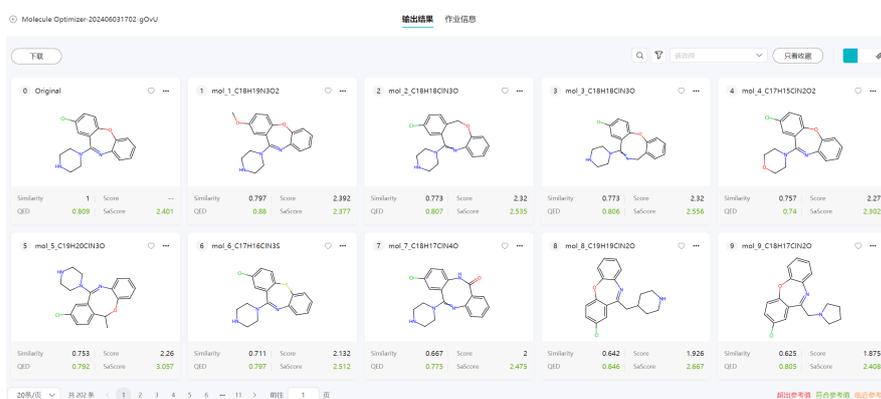
图 1-105 排序方式



- 单击每个分子卡片右上方的 可以收藏结果，收藏后可直接在收藏夹页查看。
- 单击每个分子卡片右上方的 ，可以选择“查看详情”、“查看3D”、“下游分析”、“下载3D”。
  - 查看详情：单击查看详情，跳转至分子详情页进行查看。
  - 查看3D：查看分子的3D视图。
  - 下游分析：分子优化对应的下游分析为分子优化、分子搜索和合成路径规划，如果分子设置了靶点，可以选择自由能微扰进行下游分析，单击“确定”即可创建。

- 下载3D: 如果分子设置了靶点, 可以单击“下载3D”下载优化后的小分子或者复合物, 如果分子未设置靶点, 单击“下载3D”只能下载小分子。小分子下载支持SDF和PDB格式, 复合物下载只支持PDB格式。
- 每个分子卡片上会展示相应分子序号与对应的参数Vina Score ( 有靶点 )、Score、Similarity、QED、SaScore
  - Vina Score: 代表分子如果添加了靶点, 将会计算对接结合能, 并按照 Vina Score进行排序
  - Score: 代表优化后小分子的综合打分。
  - Similarity: 代表优化后的分子与原始分子的相似度。
  - QED: 代表分子的成药性。
  - SaScore: 代表合成可及性分数, 旨在评估分子的合成难易程度。

图 1-106 查看结果 ( 2 )



单击“查看3D”, 可以看到分子的3D构象, 如果设置了靶点, 还可以看到优化后的小分子与靶点的结合构象。

如果上传了双靶点, 可以通过切换 < 1 / 2 > 来切换靶点, 查看相应靶点和优化后分子的结合构象。如果设置了两个靶点会默认下载两个靶点的结果。

单击配体列表的 可以收藏结果, 收藏后可直接在收藏夹页查看。

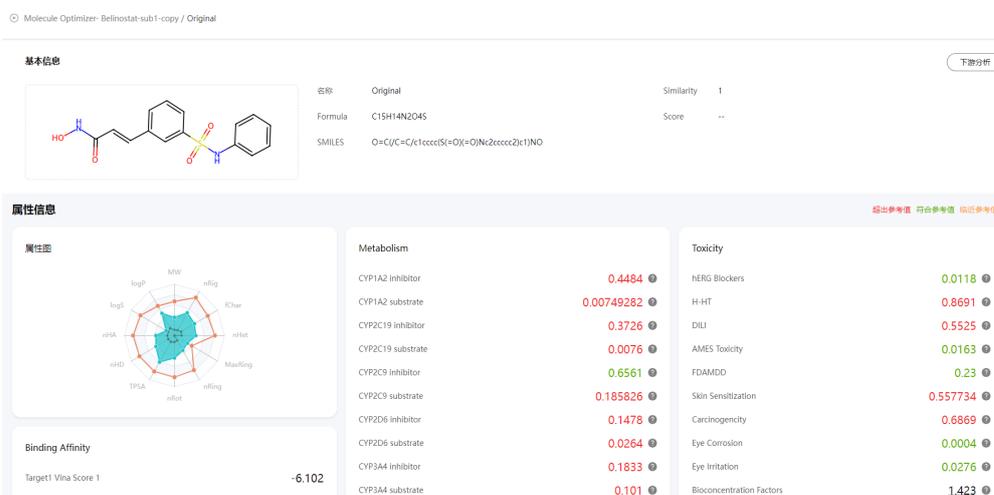
图 1-107 查看 3D 图



在查看3D的页面中，单击右侧的配体列表中，每个配体卡片右上角的 ，可以查看

- 查看详情：可以查看每个分子的属性信息，score和similarity。
- 查看属性：查看每个配体的分子属性。
- 查看2D相互作用图：查看配体中分子之间的2D相互作用图，并且可以进行图片下载。
- 下游分析：分子优化对应的下游分析为分子搜索、分子优化和合成路径，如果分子设置了靶点，可以选择自由能微扰进行下游分析，单击“确定”即可创建。
- 下载3D：选择下载小分子或者复合物后，单击“确定”，小分子支持SDF和PDB格式，复合物只支持PDB格式。

## 9. 查看分子详情



**基本信息**

名称 Original Similarity 1  
Formula C15H14N2O4S Score --  
SMILES O=C(C=C(C(=O)Nc1ccc(S(=O)(=O)Nc2ccccc2)cc1)NO

**属性信息**

**属性图**

Binding Affinity  
Target1 Vina Score 1 -6.102

**Metabolism**

CYP1A2 inhibitor	0.4484
CYP1A2 substrate	0.00749282
CYP2C19 inhibitor	0.3726
CYP2C19 substrate	0.0076
CYP2C9 inhibitor	0.6561
CYP2C9 substrate	0.185826
CYP2D6 inhibitor	0.1478
CYP2D6 substrate	0.0264
CYP3A4 inhibitor	0.1833
CYP3A4 substrate	0.101

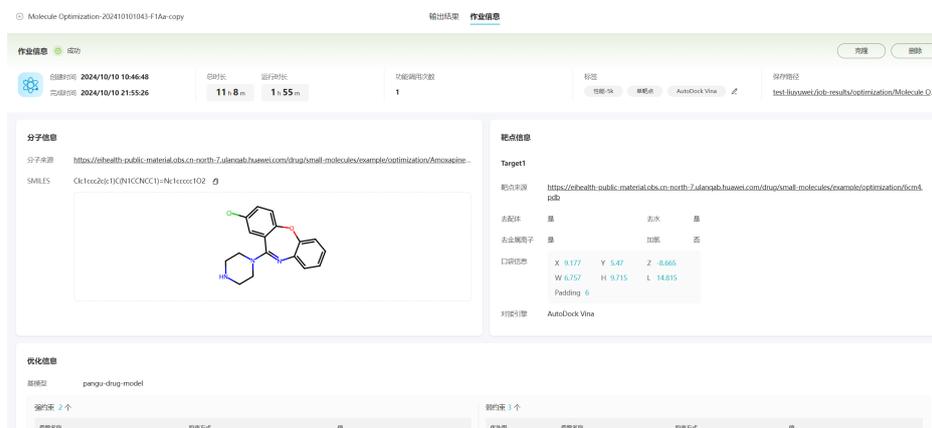
**Toxicity**

hERG Blockers	0.0118
H-HT	0.8691
DILI	0.5525
AMES Toxicity	0.0163
FDAMGD	0.23
Skin Sensitization	0.557734
Carcinogenicity	0.6869
Eye Corrosion	0.0004
Eye Irritation	0.0276
Bioconcentration Factors	1.423

## 10. 查看作业信息

单击“作业信息”切换到作业信息页签，可以查看作业的分子信息、靶点信息与优化信息等。

图 1-108 作业信息



**作业信息**

创建时间 2024/10/10 10:46:48  
完成时间 2024/10/10 21:55:26  
评价 11.8 m  
运行时长 1.55 m  
可编辑次数 1

**分子信息**

分子来源 <https://ehealth-public-material.obs.cn-north-7.aliyuncs.com/ctug/small-molecules/example/optimization/arrowajine...>  
SMILES Ck1ccc2c(c1)nc(NC(=O)N)nc2

**靶点信息**

靶点来源 <https://ehealth-public-material.obs.cn-north-7.aliyuncs.com/ctug/small-molecules/example/optimization/5cm4.pdb>

名称	类型	是否
去配体	是	否
去金属离子	是	否

对接坐标  
X: 9.177 Y: 5.47 Z: -8.665  
W: 6.757 H: 9.715 L: 14.815  
Padding: 6  
对接引擎: AutoDock Vina

**优化信息**

模型名 pangou-drug-model  
策略数 2 个  
策略名 1 个

## 聚类分析

目前分子优化返回的结果小分子数较多，无法进行批量分析，通过一些聚类的辅助方式能更好的选择分子。从每个类里挑选出一两个分子进行后续分析和验证，提高分析

的效率和分析质量。也可以通过聚类找出一些关键的骨架，来进行下游分析或者优化等。

1. 在输出结果页面左上角单击“聚类分析”后，系统开始进行分析，同时显示“聚类分析中”。

图 1-109 聚类分析



2. 待聚类分析完成后，单击“查看聚类结果”。进入聚类结果页。
3. 在聚类结果页面，可以查看每个聚类的分子数量等信息。

图 1-110 查看聚类结果

聚类方法 HIQ MC	分子数量 149						
下载							
序号	公共子结构	公共子结构原子数	分子数量	SMILES	平均Tina Score	平均Score	操作
1		21	7	C1=CN=CC(=CN)C3=NC=CC=C1C...	-6.501	0.300	查看详情
2		21	3	C1=CN=CC(=CN)C3=NC=CC=C1C...	-6.501	0.300	查看详情
3		21	5	C1=CN=CC(=CN)C3=NC=CC=C1C...	-6.501	0.300	查看详情
4		21	4	C1=CN=CC(=CN)C3=NC=CC=C1C...	-6.501	0.300	查看详情
5		21	5	C1=CN=CC(=CN)C3=NC=CC=C1C...	-6.501	0.300	查看详情

4. 单击某个聚类的操作列的“查看详情”，即可进入聚类详情页面，聚类详情页支持以卡片、列表以及3D的形式查看。默认展示卡片页面，用户可自行进行切换。

### 说明

- 每个结果页面只用进行一次聚类分析操作。
- 聚类结果是存成文件，如果文件被删或者获取不到的话会有警告，聚类结果不存在。此时可以单击“重新聚类分析”。
- 如果聚类失败，根据提示失败原因解决问题后，可单击“重新聚类分析”。

## 相关参数

分子优化的具体参数，请参考[相关参数](#)。

### 1.7.3.2 靶点口袋分子设计

基于盘古药物分子大模型，靶点口袋分子设计功能主要是能够根据给定的口袋和小分子利用AI的预测出更优小分子。

1. 单击“靶点口袋分子设计”功能卡片，进入配置页面。
2. 在配置页面上选择设计方式
  - a. 设计方式：支持侧链修饰、骨架跃迁、片段生长和从头生成四种方式。

- 侧链修饰：会在不同的侧链上进行侧链生长，生成新颖小分子。
  - 骨架跃迁：会替换您所去掉的关键骨架，生成新颖的小分子。
  - 片段生长：选择保留部分片段后，会在您删除的方向进行补充，生成新颖的小分子。
  - 从头生成：通过指定的口袋从头生成小分子。
3. 选定设计方式后，上传靶点结构，并配置其他参数。

图 1-111 查看配置页面



- 靶点结构：选择靶点结构文件，仅支持PDB格式。若文件中包含多个受体，则默认只处理第一个。
4. 单击下一步，配置配体文件，支持选择原始配体与上传配体文件两种方式
- 选择原始配体：选择靶点结构文件中已存在的原始配体用来定位口袋或作为片段优化，存在多个原始配体时可通过下拉框选择，与上传配体文件二选一。
  - 上传配体文件：通过上传与靶点结构有良好结合构象的配体文件确定配体，若文件中包含多个配体，默认只处理第一个，与选择原始配体二选一。仅支持SDF、MOL2、PDB格式文件，且只支持3D结构。
  - 标记位点：3D可视化区域选择位点，再单击右侧标记位点的“+”添加。最多可添加50个生长位点。
5. 单击下一步，配置其余参数
- 分子量：自适应和手动输入。自适应模式下平台根据药物分子的分子量分布进行分子设计。手动输入模式下平台会将用户输入的分子量作为强约束对分子设计过程进行限制，不合适的分子量范围可能会导致没有结果输出。
  - 选择属性模型：选择AI模型。如果需要创建模型，可参考AI模型。此参数只有专业版支持。一次最多可以选10个模型属性。
  - 名称：可修改，修改后左上角也同步修改。长度为5~64个字符；仅可以使用字母、数字、下划线“\_”、中划线“-”和空格；首位只能以数字或字母开头。
  - 标签：设置任务标签。
  - 功能调用消耗：运行成功后功能会消耗一次。
6. 单击“提交”。

提交成功后，可以在“作业中心”查看执行结果。

7. 查看运行结果。

- 作业执行完毕后，单击作业名称，可进入输出结果页面。序号为0的分子为原始分子，在从头生成设计模式下不存在序号为0的分子。
- 生成的分子按照与靶点结构结合度评分Score从小到大进行排序，用户也可在此页面直观查看各分子的多种药化性质。
- 结果页面支持对生成分子进行查看详情、查看3D结构、下游分析（分子优化及合成路径分析）、与原始分子属性对比（仅支持片段优化的属性对比）、查看分子与靶点的2D相互作用、下载2D相互作用图片，下载3D分子结构、收藏结果等多项功能。
- 支持按照相互作用力进行高级筛选，单击进行条件配置。
- 可以以列表的形式查看口袋分子的作业，单击“下载”，下载分子的基本信息和属性或者分子3D构象。

 说明

下载操作将产生流量费用，具体可参考[计费说明](#)。

8. 靶点口袋分子设计结果支持以卡片视图的形式进行查看，参考图1-115，在卡片视图中：

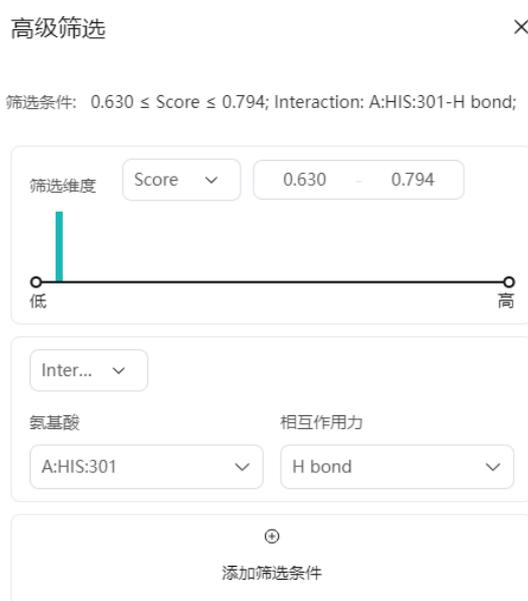
- 单击  可以收藏结果，收藏后可直接在收藏夹页查看。
- 单击右上方的选择下拉框，可以选择分子的排序方式，将分子按照所选的排序方式进行展示。

图 1-112 排序方式



- 单击右上方的筛选，可以进行属性和相互作用力筛选。

图 1-113 高级筛选



- 单击每个分子卡片右上方的 **☰**，可以选择“查看详情”、“查看3D”、“下游分析”、“下载3D”。
  - 查看详情：单击查看详情，跳转至分子详情页进行查看。
  - 查看3D：查看分子的3D视图。
  - 下游分析：靶点口袋分子设计对应的下游分析为分子优化、自由能微扰、合成路径规划和分子搜索，单击“确定”即可创建。
  - 下载3D：如果分子设置了靶点，可以单击“下载3D”下载优化后的小分子或者复合物，如果分子未设置靶点，单击“下载3D”只能下载小分子。小分子下载支持SDF和PDB格式，复合物支持PDB格式。
- 每个分子卡片上会展示相应分子序号与对应的参数Score、Vina Score、Similarity (侧链修饰/骨架跃迁场景)、QED。
  - Score：代表小分子可合成性的综合打分。
  - Vina Score：代表分子如果添加了靶点，将会计算对接结合能。
  - Similarity：靶点口袋分子设计后的分子与原始分子的相似度。
  - QED：代表分子的成药性。

图 1-114 查看运行结果 (1)

序号	分子名	Formula	SMILES	Score	Vina Score	test-htgpt	Molecular Weight (g/mol)	logP	操作
1	mol_1_C19...	C19H30N2O3	CCOC(C)CNC(=O)...	...	-5.236	1.913	334.46	2.238	查看详情 查看3D 下游分析 下载3D
2	mol_2_C19...	C19H24O	C1=CC=C(C(CCC)...	...	-5.025	4.514	268.4	4.558	查看详情 查看3D 下游分析 下载3D
3	mol_3_C19...	C19H32O3	C[C@H](O)C(CCC)...	...	-4.986	2.354	308.46	3.464	查看详情 查看3D 下游分析 下载3D
4	mol_4_C17...	C17H31N5O3	CCC(C)(O)C(C)C...	...	-4.366	-0.517	353.47	-1.244	查看详情 查看3D 下游分析 下载3D
5	mol_5_C16...	C16H19F2N3O	CN(C=C2C(C)C...	...	-4.291	0.723	307.34	1.484	查看详情 查看3D 下游分析 下载3D
6	mol_6_C24...	C24H33N	NC1C2C(C)C(C)C...	...	-4.29	5.363	335.54	5.36	查看详情 查看3D 下游分析 下载3D

图 1-115 查看运行结果 (2)

分子名	Score	Vina Score	QED	SaScore
mol_1_C19H30N2O3	...	-5.236	0.799	5.347
mol_2_C19H24O	...	-5.025	0.626	4.637
mol_3_C19H32O3	...	-4.986	0.821	4.021
mol_4_C17H31N5O3	...	-4.366	0.482	5.251
mol_5_C16H19F2N3O	...	-4.291	0.745	5.283
mol_6_C24H33N	...	-4.29	0.734	4.998
mol_7_C15H21F	...	-4.27	0.535	5.414
mol_8_C13H18N2O2	...	-4.186	0.682	5.162
mol_9_C20H33N	...	-4.118	0.698	4.221
mol_10_C13H16F5N5O5	...	-4.089	0.654	4.921



1. 在输出结果页面左上角单击“聚类分析”后，系统开始进行分析，同时显示“聚类分析中”。

图 1-119 聚类分析



2. 待聚类分析完成后，单击“查看聚类结果”。进入聚类结果页。

图 1-120 查看聚类结果

聚类号	公共子结构	公共子结构原子数	分子数	SMILES	平均Vina Score	操作
1		24	2	NC1=NN=C(C(=C)C=NC3=NC=C(C4=CNC5=CC=C...	-8.034	查看详情
2		23	81	C1=CC2=C(C(=C)CN=C4NC=C(C(=NN=COS)C4#...	-8.435	查看详情
3		23	51	C1=NC2=NC=C(C(=C)NN=CO3)C2C=C1C1=CNC...	-10.731	查看详情
4		23	49	[C1]=[C][C]=C2C(C3=[C]=C4N[C]=C(C5=NN=...	-8.841	查看详情
5		23	49	C1=CC2=C(C(=C)CC4C(C5=NN=COS)=CN=C4N=...	-10.558	查看详情

3. 在聚类结果页面，可以查看每个聚类的分子数量等信息。
4. 单击某个聚类的操作列的“查看详情”，即可进入聚类详情页面，聚类详情页支持以卡片、列表以及3D的形式查看。默认展示卡片页面，用户可自行进行切换。

图 1-121 查看聚类详情

## 📖 说明

- 每个结果页面只用进行一次聚类分析操作。
- 聚类结果是存成文件，如果文件被删或者获取不到的话会有警告，聚类结果不存在。此时可以单击“重新聚类分析”。
- 如果聚类失败，根据提示失败原因解决问题后，可单击“重新聚类分析”

## 相关参数

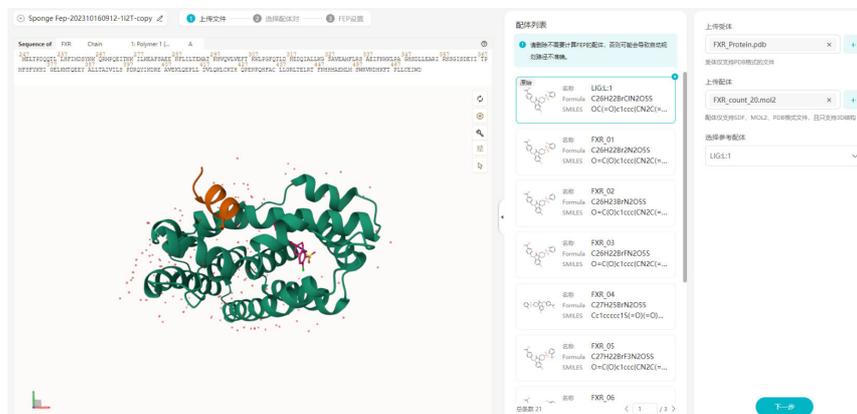
靶点口袋分子设计的具体参数请参考[相关参数](#)。

### 1.7.3.3 自由能微扰

自由能微扰基于纯国产分子动力学模拟库SPONGE，产生自动化FEP workflow，端到端计算配体修饰造成的亲和能改变。

1. 单击“自由能微扰”功能卡片，进入上传文件页面。
2. 在上传页面右侧，选择上传受体，上传配体，选择中心配体。
  - 上传受体：受体仅支持PDB格式的文件。
  - 上传配体：配体仅支持SDF、MOL2、PDB格式文件，且只支持3D结构。
  - 选择参考配体：当前自由能微扰支持自动规划路径，选择参考配体后系统自动计算，用户也可自主添加或删除配体对。

图 1-122 上传文件



## 📖 说明

引用外部桶时，需要确保所引用的数据不超过45层级的目录。

3. 单击“下一步”，选择配体对。
 

页面显示：正在规划自动路径，您也可以直接选择配体对后进行下一步。

待计算路径：选择待计算的路径。待计算路径起点是中心配体名称，终点是其他配体的名称。在相似度计算完成之前默认未勾选。您也可以添加路径或者重置路径。添加路径和重置路径可以通过单击右边的“添加路径”或者“重置”进行操作。添加路径也可以在左侧微扰图中直接通过两个分子之间进行连线添加。可以在微扰图中单击某条待计算路径上的 ，删除该条待计算路径。

图 1-123 添加或者删除待计算路径

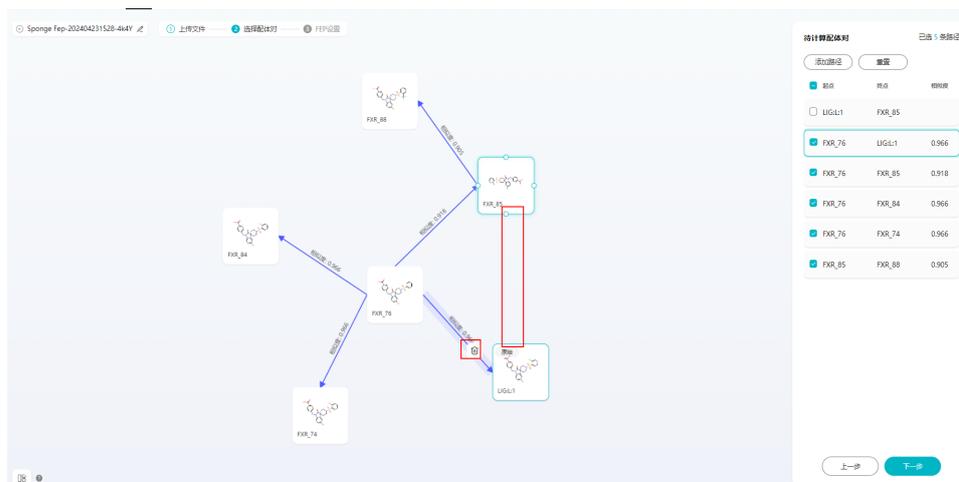
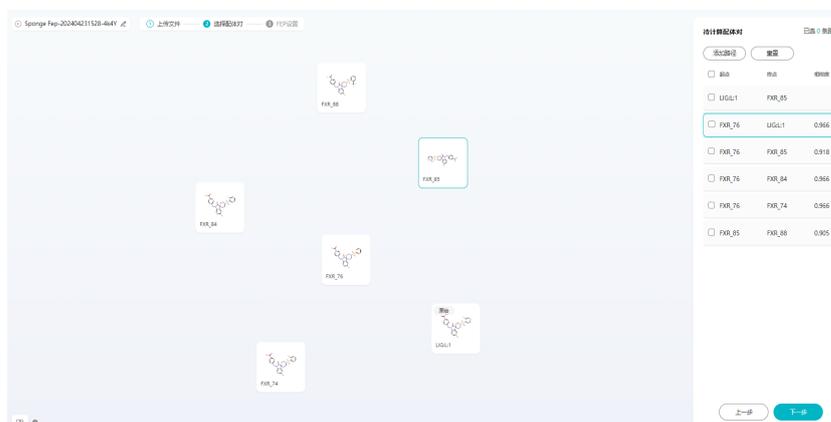
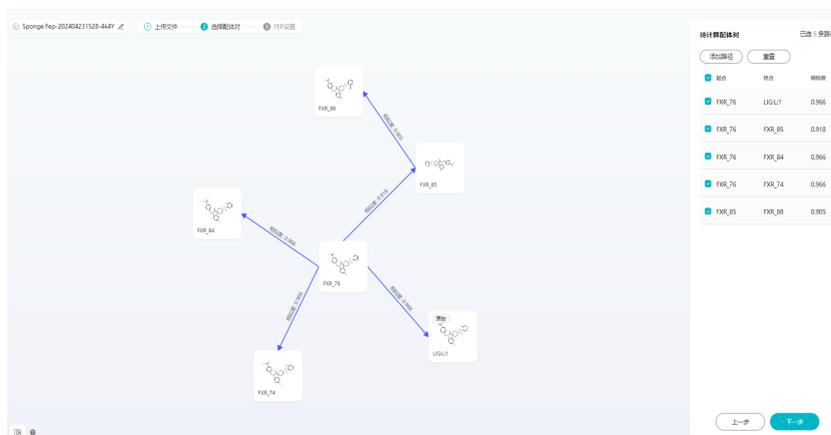


图 1-124 选择配体对



返回相似度后默认全勾选，您可以进行勾选或去除勾选要计算的路径，如果未勾选，则后面就不会对其进行FEP计算。在相似度返回之前，您也可以直接选择配体对进入下一步。

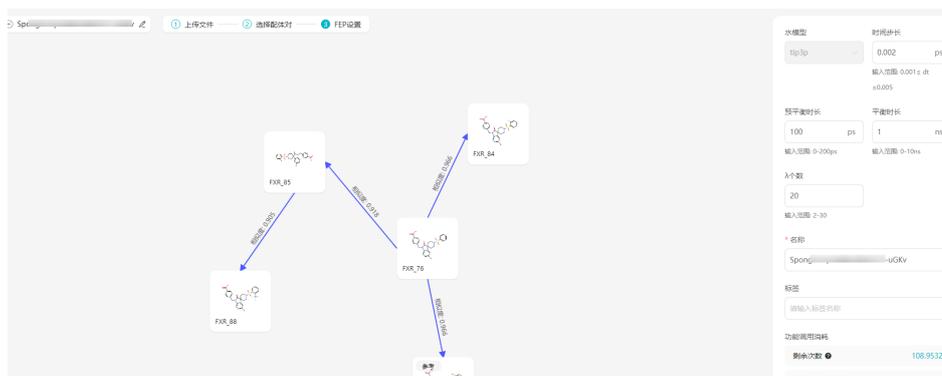
图 1-125 选择计算路径



4. 单击“下一步”，进入FEP设置页面，设置相关参数。

- 时间步长：默认值：0.002，取值范围： $0.001 \leq dt \leq 0.005$ ，单位：ps。分子动力学模拟的步长，建议不超过0.002ps，步长越大，越难收敛。
- 预平衡时长：默认值：100ps，取值范围：0-200ps。对体系进行预平衡模拟，使体系温度、压强、密度等达到平衡状态。预平衡模拟时长=预平衡步数×时间步长。时长增加，作业运行时间延长。
- 平衡时长：默认值：1ns，取值范围：0-10ns。平衡阶段的模拟，用于自由能微扰计算。平衡模拟时长=时间步长×平衡步数/1000，单位为ns。时长增加，作业运行时间延长。
- λ个数：默认值20，输入范围为2-30。自由能微扰的窗口数量。
- 名称：可修改，修改后左上角也同步修改。长度为5~64个字符；仅可以使用字母、数字、下划线“\_”、中划线“-”和空格；首位只能以数字或字母开头。
- 标签：设置任务标签。
- 功能调用消耗：每一对会消耗一次功能调用，因此计算几条路线就显示调用几次。

图 1-126 设置 FEP 参数



5. 单击“提交”，可在作业中心查看该作业的运行情况。
6. 查看运行结果。
  - 输出每对配体的相对结合自由能、分子图、相似性等。也可以单击右边“查看轨迹”，下载运动轨迹。结果页面支持Pair和Ligand两种查看方式。Ligand查看方式下，可对结果进行收藏，在Ligand中的收藏同步3D视图。如果取消收藏，单击🏠，弹出取消收藏页面，单击“确认”，取消收藏。
  - 也可以下载输出结果文件包含小分子的基本信息和属性。

### 📖 说明

下载操作将会产生流量费用，具体可参考[计费说明](#)。

图 1-127 查看结果 (1)

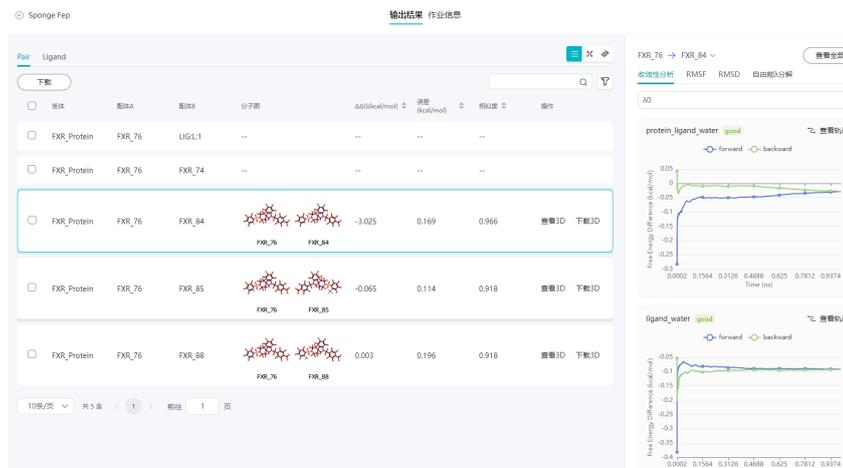
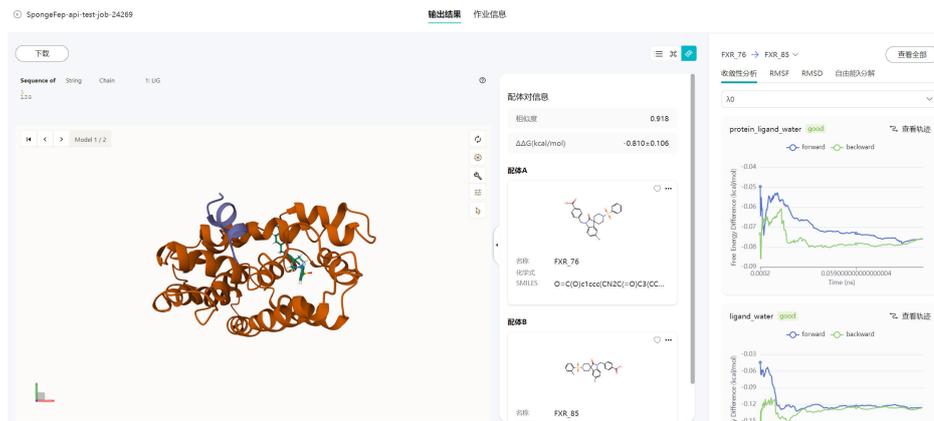


图 1-128 查看结果 (2)

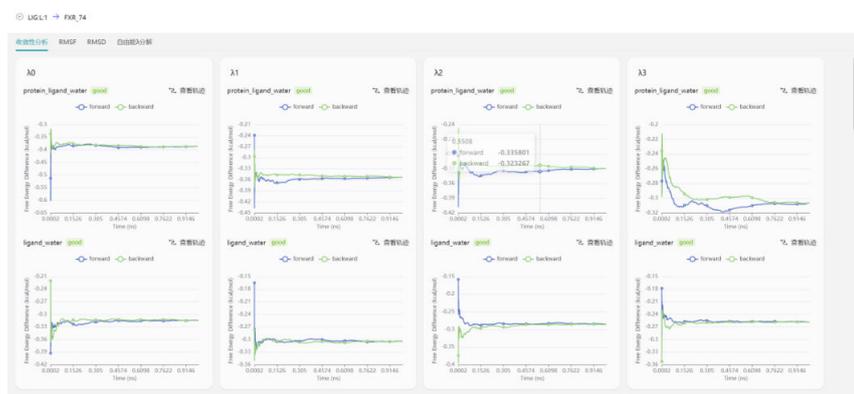


图 1-129 查看结果 (3)



- 单击**查看全部**展示收敛性分析、RMSF和RMSD结果。

图 1-130 查看结果 (4)



- 单击查看轨迹预览轨迹动图。

图 1-131 查看结果 (5)



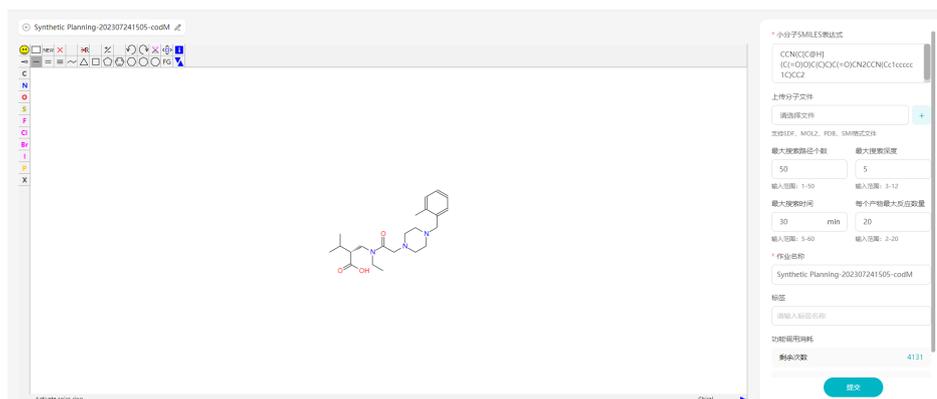
### 1.7.3.4 合成路径规划

合成路径规划基于盘古药物分子大模型，根据给定的目标分子，可以设计出完整且合理的合成路径。

1. 单击“合成路径规划”功能卡片，进入配置页面。
2. 在配置页面，可以在左侧绘制分子，也可以通过上传分子文件方式上传分子或者在白框内输入小分子SMILES表达式。
  - 上传分子文件：支持SDF、MOL2、PDB、SMI格式文件。
  - 最大搜索路径个数：合成路径规划的路径数量。路径数量增加，将展示更多的合理合成路径；路径数量减少，可能会有部分合理路径未展示。默认值50，取值范围1-50。
  - 最大搜索深度：深度增加，每一个路径可进行搜索的深度限制增加，作业运行时间可能延长；深度减少，部分路径可能在还未搜索完成时被终止。默认值5，取值范围3-12。
  - 最大搜索时间：合成路径规划的搜索时间限制，到达限制时间会返回已经搜索完成的路径。搜索时间增加，作业运行时间延长；搜索时间减少，可能会有部分合理路径未能开始搜索。默认值30min，取值范围5-60。
  - 每个产物最大反应数量：合成路径中每一个中间产物最多有多少种可能的反应，按照反应置信度从高到低排序。最大反应数量增加，每个中间产物的可搜索反应范围数量增加，作业运行时间延长；最大反应数量减少，可能会有部分合理反应未能纳入搜索。默认20，取值范围2-20。

- 作业名称：可修改，修改后左上角也同步修改。长度为5~64个字符；仅可以使用字母、数字、下划线“\_”、中划线“-”和空格；首位只能以数字或字母开头。
- 标签：设置作业标签。
- 功能调用次数：合成路径规划目前是一个运行成功得作业消耗一次功能调用次数。

图 1-132 分子合成路径

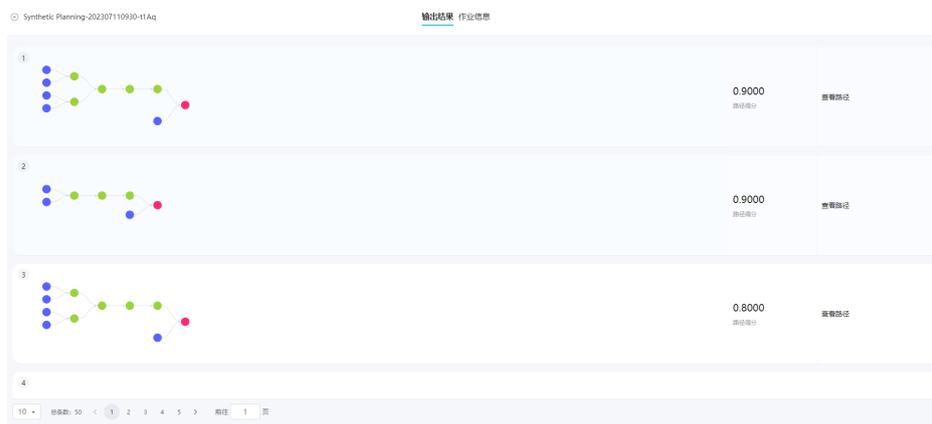


### 说明

引用外部桶时，需要确保所引用的数据不超过45层级的目录。

3. 单击“提交”。可在作业中心查看该作业的运行情况。  
运行完成后，可在作业中心单击该作业查看输出结果，输出结果缩略图。

图 1-133 查看运行结果（1）



4. 单击查看路径，查看输出结果详情。  
可以单击左上角“下载”，下载当前的输出结果。

### 说明

下载操作将会产生流量费用，具体可参考[计费说明](#)。

图 1-134 查看运行结果（2）



## 1.7.4 通用工具

### 1.7.4.1 MOL 3D Viewer

Mol 3D Viewer工具支持查看蛋白质和小分子的3D结构。单击“功能模块 > 通用工具 > MOL 3D Viewer ”，进入MOL 3D Viewer页面即可操作。

主要功能包括：

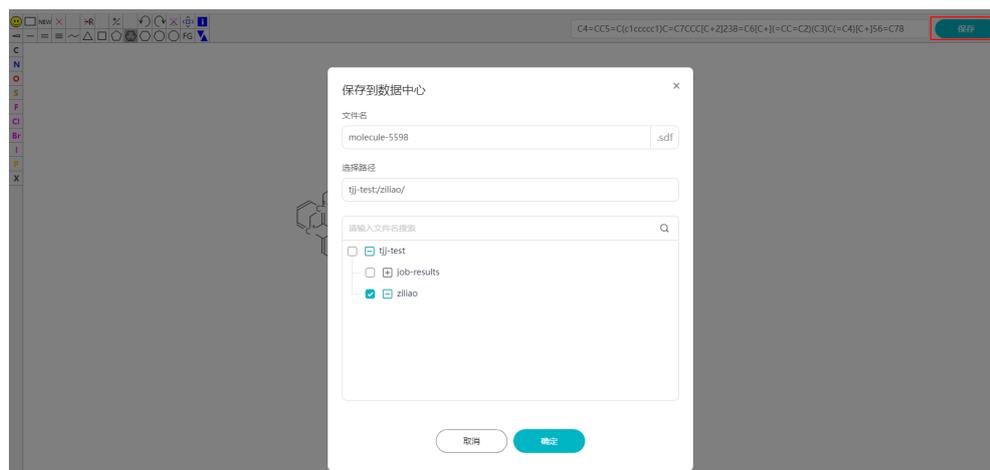
- 本地分子结构的3D可视化，支持多种显示样式，支持本地文件拖拽进画布区域进行渲染。
- 大分子支持PDB编号自动下载。
- 相互作用的可视化。
- 支持分子的编辑、标注。
- 测量、测量距离、角度、二角面等。
- 结构叠合，将不同结构叠合在一起，用于比较结构差异。
- 输出指定蛋白口袋的位置及大小，用于分子对接。
- 导出不同分辨率的图片文件。
- 播放与制作分子动力学轨迹动画。

### 1.7.4.2 MOL Editor

Mol Editor工具可以查看小分子的2D结构，以及编辑分子结构。

单击“功能模块 > 通用工具 > MOL Editor ”，进入Mol Editor页面即可操作。可以通过单击右上角的“保存”，将编辑的分子结构保存在数据中心。

图 1-135 保存分子结构

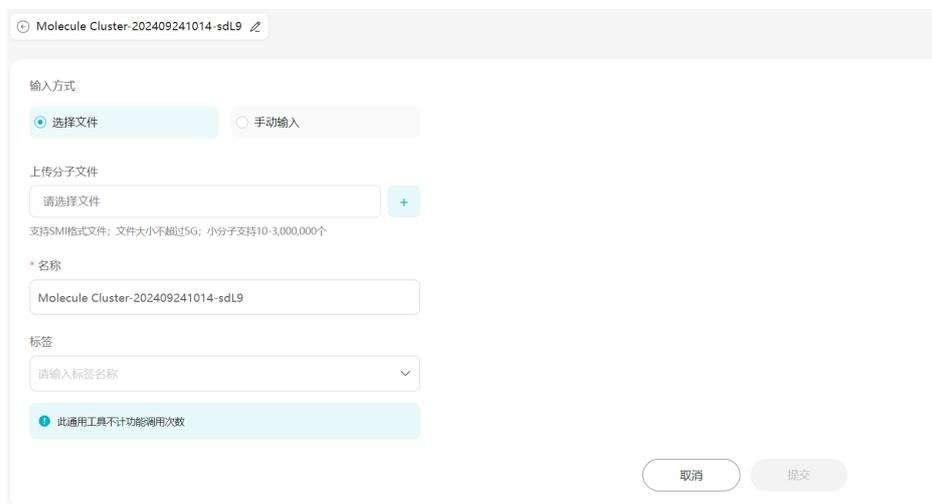


### 1.7.4.3 聚类分析

聚类分析工具可以通过骨架聚类方法，将大型小分子数据库中结构相似的化合物聚成一类，从而找到有效骨架，辅助苗头化合物发现。

1. 单击“功能模块 > 通用工具 > 聚类分析”功能卡片，进入配置页面。

图 1-136 聚类分析配置页面



- 输入方式：选择文件和手动输入类型。
  - 上传分子文件：选择可上传的分子文件。分子文件支持SMI格式文件；文件大小不超过5G；小分子支持10-10,000个，超出10,000的分子会进行截断。
  - 输入小分子：可以通过上传文件或输入SMILES以输入小分子。
  - 名称：设置作业名称。长度为5~64个字符，仅可以使用字母、数字、下划线“\_”、中划线“-”和空格，首位只能以数字或字母开头。
  - 标签：设置任务标签。
2. 参数设置后单击“提交”。可在作业中心查看该作业的运行情况。

运行完成后，可在作业中心单击该作业查看输出结果。

- 查看运行结果。在聚类结果页面，可以查看每个聚类的公共子结构、公共子结构原子数、分子数量等信息。

可以单击“下载”，下载聚类分析结果信息。下载结果以Excel格式输出，包含小分子基本信息和属性信息。

图 1-137 查看结果

序号	公共子结构	公共子结构原子数	分子数	SMILES	平均得分	操作
1		32	10	C1-CC-[N+][C@H]2CC[C@@H]3COPPOC[C@H]2C1	0.69463	查看详情
2		30	7	C(CCC1-CC-CC-C1)N[C@@H]1CN(CCC-CC-C)C1	0.68546	查看详情
3		30	3	C(NC1CN2CCN(CCC[C@@H]3C2O1)C[C@@H]1CN)C2	0.7201	查看详情
4		30	3	C(N1CN(CCN1)N(C)C2-CC-CC-C2)-N[C@@H]1C	0.70439	查看详情
5		30	3	C(N[C@@H]1)CN(CCC[C@@H]1)CC1-CC-C(C)C2-C1	0.77962	查看详情
6		30	3	C1C-C1-C-C-C-C-C-C-C-C-C-C-C-C-C-C-C-C	0.71537	查看详情

- 单击操作列的“查看详情”，可查看聚类详细信息。聚类详情页支持以卡片、列表的形式查看，默认展示卡片页面，用户可自行切换。

支持收藏功能，单击按钮，即可收藏CPI预测结果，收藏后可直接在收藏夹页查看。

图 1-138 查看聚类详细信息

公共子结构

Similarity 分布

- 查看作业信息页面。从作业结果页面单击“作业信息”查看作业信息界面。

图 1-139 查看作业信息

作业信息

创建时间: 2024/09/24 09:53:51 | 完成时间: 2024/09/24 09:58:45

耗时: 4 = 54 | 运行时长: 3 = 43

功能调用次数: 0

保存路径: jingqi/job.results/clustering/Molecule Cluste...

小分子信息

小分子个数: 8775 / 8777

小分子来源: https://ehealth-public-material.obs.cn-north-2.huaweicloud.com/drug/small\_molecules/public/drugbank\_8777.xml

## 1.8 AI 模型

### 创建模型

盘古辅助制药平台支持用户创建AI模型，目前AI模型只有专业版支持。AI建模支持创建属性模型和基模型。创建属性模型是基于自定义数据，对盘古药物分子大模型进行微调，进行属性预测和迭代活性优化，实现干湿实验闭环。基模型基于自定义化合物数据，对盘古药物分子大模型进行增量预训练，提升化合物表征精度。

1. 登录盘古辅助制药平台，选择“AI模型”。

图 1-140 AI 模型



2. 单击“创建模型”，设置相关参数信息。

表 1-6 参数说明

参数	说明
名称	模型名称。 长度为5-32个字符，首位需以英文字母开头，仅可以使用字母、数字、下划线“_”和中划线“-”和空格。
描述	设置模型的描述信息。
基模型	设置基模型，基模型与分子描述符有关。选择不同的基模型，分子描述符不一样。
模型类型	<ul style="list-style-type: none"> <li>● 回归型：预测一系列连续变量的模型，主要侧重定量描述。</li> <li>● 二分类：预测二分类离散变量的模型，主要侧重定性分析。</li> </ul>
模型数据	选择模型数据。可选择数据中心数据，或者示例数据。 仅支持CSV格式，数据条数支持100-50000条，文件不大于5MB。
组织共享	如果关闭，则该模型只能自己使用。 如果开启组织共享，则其他子用户也可以使用您的模型做分析。 默认为关闭。

3. 单击“确定”。

每人最多可以创建100个模型，每次使用模型时，最多可以使用10个。

## 查看模型列表

在AI模型页签下支持查看创建的所有模型。包括模型的名称、模型类型、基模型来源、创建时间、完成时间、创建者、状态、组织共享、操作等信息。

图 1-141 查看属性模型列表

名称	模型类型	基模型来源	创建时间	完成时间	创建者	状态	组织共享	操作
test-model-34760	二分类	pangu-drug-model	2024/06/08 15:17:19	2024/06/10 17:21:22		可用		查看Loss
test-model-12992	二分类	pangu-drug-model	2024/06/03 14:33:25	2024/06/03 15:49:19		可用		查看Loss
erfensixing	二分类	tasly-custom-base-m...	2024/05/30 10:06:55	2024/05/30 19:24:45		可用		查看Loss
huiguxing	回归值	pangu-drug-model	2024/05/30 10:06:19	2024/05/30 14:07:12		可用		查看Loss

- 查看模型指标

- 查看loss值：loss代表模型训练的损失变化。  
单击相应模型操作列的“查看loss”即可查看相应的训练集Loss。
- 删除模型：单击相应模型操作列的“删除”，在弹窗中单击“确定”，即可删除掉对应的模型。
- 查看评价指标：在模型列表页，单击某个模型名称左侧的 > 按钮，可展示当前模型的相关指标，包括模型的数据量、描述、区间范围、评价指标、模型数据。  
其中，评价指标的值代表了训练完成的模型在测试集上的好坏。

## 查看基模型

单击“基模型”，可以查看基模型列表。内置官方盘古药物大模型，在公测期间限时免费。

基模型名称	创建时间	描述
pangu-drug-model	2024/05/20 11:31:09	官方预置模型
t...	2024/05/27 11:17:15	

## 1.9 自定义数据库

盘古辅助制药平台支持用户自定义数据库，可以上传自己的数据构建自己的数据库，进行后续的分子搜索。构建自定义数据库需要先购买CSS资源和绑定CSS资源，详情见[资源中心](#)

### 自定义数据库

- 返回主页，单击“自定义数据库”进入界面，单击页面左上角的“创建数据库”，并填写相关信息。每个用户可新建100个自定义数据库。

图 1-142 创建数据库

- 名称：数据库名称。长度为5-32个字符，首位需小写英文字母开头，仅可以使用小写字母、数字、下划线“\_”和中划线“-”。最终租户下的所有用户数据库名称不能重名，可以和官方数据库名称重名，不区分大小写。目前支持的官方数据库有DrugBank、ZINC、TopSciense、DrugSpaceX。
- 描述：输入数据库描述，支持中英文，上限为1024个字符。
- CSS集群名称：选择已经绑定好的CSS集群，将会在这个集群上构建数据库。
- 数据文件：输入数据库文件，仅支持CSV格式，文件不大于1G，数据条数支持最多5,000,000条，超过5,000,000条将进行截取。文件必须有两列是“SMILES”、“NAME”，不区分大小写。文件不允许有中文，数据库小分子会自动去重。
- 数据文件列名：选择数据库列名，仅解析前100个列名，支持选择12个列名保存到数据库。
- 组织共享：
  - 如果关闭，则该数据库只能自己使用。
  - 如果开启组织共享，则其他子用户也可以使用您的数据库进行分子搜索。
  - 默认为关闭。
- 数据库列表展示数据库的名称、分子数量、创建时间、状态等信息。对于已创建的数据库，可以执行删除和追加数据操作。数据库状态介绍如下：
  - 可用：数据库正常，可以进行分子搜索使用。

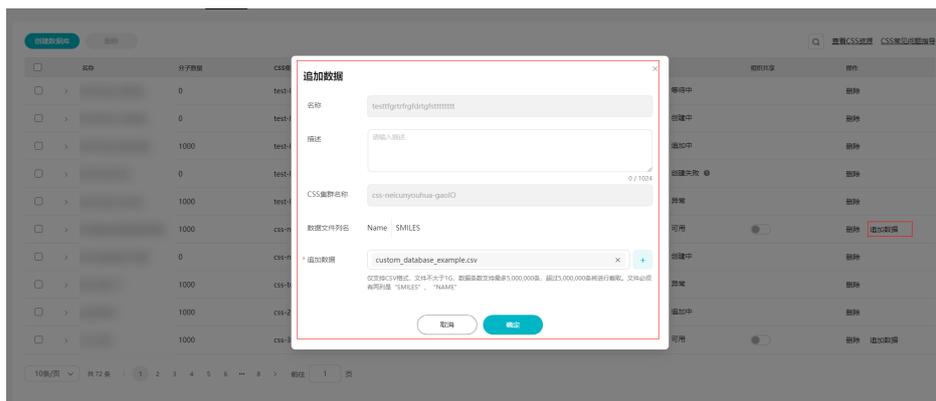
- 等待中：数据库创建任务正在等待。
- 创建失败：数据库创建失败。
- 创建中：数据库正在创建。
- 追加中：数据库正在追加，此时数据库可以正常使用，搜索的数据库是正在追加过程的数据库。
- 异常：数据库异常，可能是由于CSS到期、CSS密码改变、CSS资源解绑等问题导致数据库异常。

图 1-143 删除数据库和追加数据



- 可以通过“追加数据”到数据库来增加数据条数。每次追加最多追加500万条数据。

图 1-144 追加数据



### 说明

CSS相关常见问题请参考常见问题文档中的CSS章节。

## 1.10 数据管理

### 1.10.1 数据管理简介

盘古辅助制药平台使用对象存储服务（OBS）存储原始数据、作业执行中间数据和执行结果数据。数据按项目维度进行隔离和划分，从项目角度进行数据的管理，不同项目的数据可以通过“复制数据”实现跨项目使用。您可以在“项目管理”页面，“数据”页签中，完成数据的添加、导入、归档、复制、删除等操作。

## 说明

- 数据总条数 $\leq 10000$ ，页面上可以指定跳转具体页码。
- 数据总条数 $> 10000$ ，通过“下一页”和“上一页”进行页面切换，并且按时间排序只能对当前页面进行排序，页面搜索为按前缀搜索。

## 使用合规数据

您添加的数据属于您的内容，您应对您的内容的合法合规性负责，建议您根据对您可适用的法律法规，进行相应的脱敏、匿名、加密或其他符合要求的处理或采取相应保护措施，当您在复制或者引用数据的时候，需审视是否存在隐私数据，请谨慎操作。

## 上传数据方式

在开始执行分析作业前，请先上传待分析的原始数据。不同的上传方法对数据大小要求不同，您可以参考[表 上传数据方式](#)选择相应的数据上传方式。

表 1-7 上传数据方式

上传数据方式	说明
“数据中心”页面上传	通过“数据中心”页面上传数据，支持上传最大为1GB，最多20个文件。 数据上传方法请参见“ <a href="#">数据</a> ” <a href="#">页面上传</a> 。

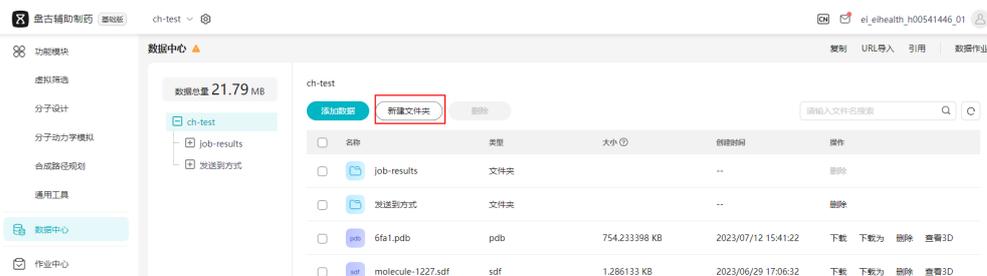
### 1.10.2 添加数据

在添加数据前，您可以创建存储数据的文件夹，将数据存储在对应的文件夹中。如果已有存储数据的文件夹，可以直接添加数据。

#### （可选）新建文件夹

1. 在项目页面选择“数据中心”。
2. 单击“新建文件夹”，创建文件夹。

图 1-145 新建文件夹







### 说明

- 引用OBS类型数据时，如果数据在OBS中的存储类型为“归档存储”，则将该数据引用过来后，该数据不能用于创建作业，并不可下载。
- 平台系统管理员在自己的所有者、管理员、操作者项目可以引用OBS类型数据。平台系统管理员在自己的所有者、管理员、操作者项目可以解除引用OBS类型数据。其他角色的用户仅能使用引用进来的OBS类型数据。

## 使用 URL 导入数据

使用URL方式将公网http/https/ftp链接的数据或者OBS中的数据导入。

1. 单击数据中心右上角“URL导入”。
2. 单击“添加URL”，填写URL地址。
3. 选择需要导入数据的文件夹。
4. 单击“确定”，导入数据。

图 1-149 使用 URL 导入



### 1.10.3 数据管理常用操作

- [复制数据](#)
- [下载数据](#)
- [删除数据](#)
- [查看3D](#)

执行数据管理类操作需要项目成员具备相应的权限，详细的权限介绍请参见[项目成员和权限](#)。

数据文件的名称，不可以含有特殊字符。如果文件名包含特殊字符，将不支持下载，可通过去除文件名中的特殊字符方式解决。

#### 查看数据作业

数据复制、数据删除等操作均为系统的异步任务，您可以在“数据作业”中查看任务的执行状态。同时，平台针对该类任务支持您在“数据中心 > 数据作业”中查看任务的执行信息。在数据作业页面，支持通过作业类型、状态、创建时间、完成时间来进行搜索。

数据作业支持取消、删除、重试操作，三种操作均支持批量操作。

**取消：**一个数据作业中，单个数据操作执行过程为原子性操作，即要么不执行，要么需要执行过程不能中断。在执行完成后，“取消”操作才能生效。

**删除：**运行中的任务不支持删除，只可删除运行结束的数据。

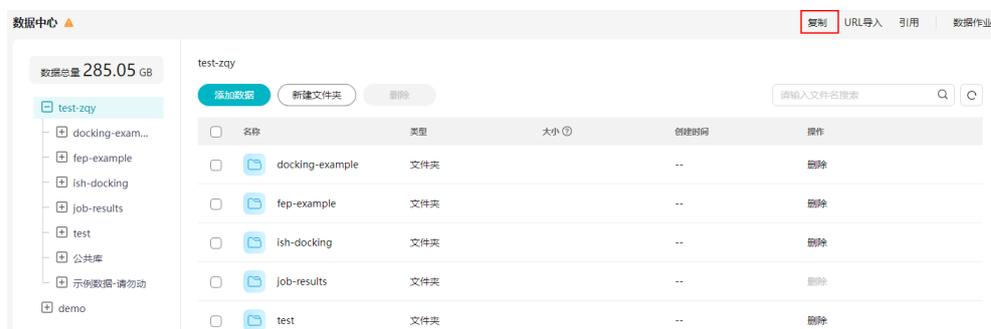
**重试：**可对执行失败的任务从失败位置进行重试。

#### 复制数据

在“数据中心”页面，使用“复制”，将本项目的某一个文件、文件夹复制到其他文件夹中。

1. 单击页面右上角“复制”。

图 1-150 复制数据



2. 选择需要复制的文件或文件夹，以及数据存储路径，单击“确认”，复制数据。

例如，将“test.zip”文件复制到“abc”文件夹中。复制数据时可单击  图标创建文件夹。

图 1-151 复制数据



## 下载数据

### 须知

下载数据操作会产生流量费用，计费方式为按需计费，计费详情请参考[OBS数据下载费用](#)。

以下操作步骤是在“数据中心”页面下载数据至本地。

1. 在“数据中心”页面，下载数据至本地。
2. 选择待下载的数据，单击“操作”列“下载”即可。  
或者单击“操作”列“下载为”，然后单击鼠标右键，选择“链接另存为”，下载数据。

图 1-152 下载数据



## 删除数据

以下操作步骤是在“数据中心”页面删除数据。

- 删除数据  
单击“操作”列的“删除”，删除数据。

图 1-153 删除数据



- 批量删除

勾选文件名左侧  图标，单击“删除”，删除数据。

图 1-154 批量删除



## 查看 3D

支持以Mol 3D Viewer工具查看蛋白质和小分子的3D结构。

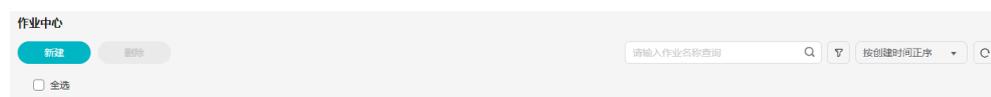
## 1.11 作业管理

### 1.11.1 作业管理简介

在作业中心页面，可以创建分子对接、分子优化、自由能微扰、合成路径规划功能的作业。

在“作业中心”页面，以列表形式展示了项目中作业的运行状态。您可以查看作业的名称、创建时间、运行状态、总时长、运行时长、已运行时间、预计还需时间。对于列表中的作业，支持通过作业名称、状态、类型、标签、创建时间和完成时间进行快速搜索。也可以根据创建时间顺序、总时间顺序、总时长顺序、完成时间顺序等进行排序。

图 1-155 作业中心



## 1.11.2 创建作业

创建作业有以下几种不同的方式，您可以任选其中一种方式。

- [通过功能模块创建](#)
- [通过作业中心页面创建](#)
  - [新建作业](#)
  - [克隆作业](#)

### 通过“功能模块”创建

1. 在“功能模块”页面，选择对应的模块，单击“立即使用”。
2. 在具体的功能模块页面，配置相关参数信息。具体参考[功能模块](#)章节。

### 通过“作业中心”页面创建

- 新建作业
  - a. 单击“新建”，在新建作业页面选择对应的模块。

图 1-156 新建作业



- b. 单击“确定”，在对应的模块页面配置详细信息，具体请参考[功能模块](#)。
- 克隆作业

克隆已有的作业，作业的参数及流程一并克隆到新作业中。

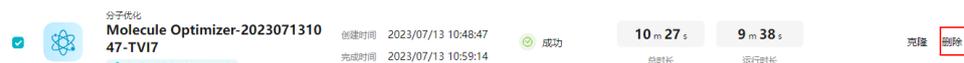
    - a. 在“作业中心”页面，单击作业操作列的“克隆”。
    - b. 在作业详情页面修改对应的参数信息。
    - c. 参数确认无误后，单击“提交”按钮。

### 1.11.3 管理作业

在“作业中心”页签，您可以查看已创建的作业。在操作列，您可以对已创建的作业执行克隆、删除操作。支持对等待中的作业进行取消操作，取消后作业不计费，运行中的作业不能取消。

- 删除单个作业  
选择需要删除的作业，单击“操作”列“删除”即可。

图 1-157 删除单个作业



- 批量删除作业  
一次选中多个作业，单击“作业中心”页面上方的“删除”。

图 1-158 批量删除作业



### 1.11.4 查看执行结果

作业运行完成后，可以在“作业中心”页面，单击作业名称，查看输出结果和作业信息。

#### 输出结果

当作业执行成功后，可以在作业中心页面，单击作业名称，查看输出结果。

在输出结果中可以查看下载某个任务或者全部任务的信息，可以在操作列查看3D、合成路径等信息。

图 1-159 输出结果-列表视图

输出结果 作业信息

序号	分子图	名称	Formula	SMILES	Score	Similarity	Molecular Weight (MW)	logP	QED	SAcore	操作
0		Original	C17H16ClN3O	Clc4ccc3oc1...	...	1	313.79	3.429	0.809	2.401	查看3D 合成路径 下载
1		mol_1_C18H...	C18H16ClN3...	O=C4C1=CC=...	2.4	0.8	341.8	2.856	0.591	2.524	查看3D 合成路径 下载
2		mol_2_C18H...	C18H19N3O2	COC1=CC=...	2.391892	0.797297	309.37	2.784	0.88	2.377	查看3D 合成路径 下载

输出结果除了支持上图所示的列表视图外，还支持卡片式图和3D视图。

图 1-160 输出结果-卡片视图

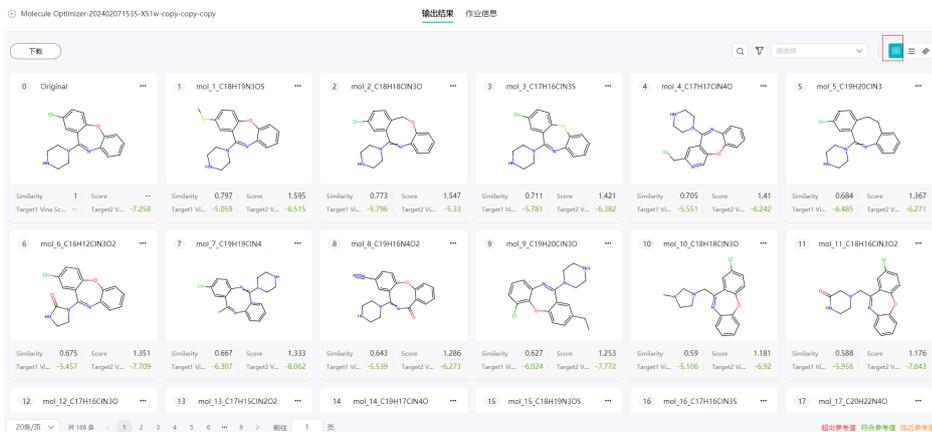


图 1-161 输出结果-3D 视图



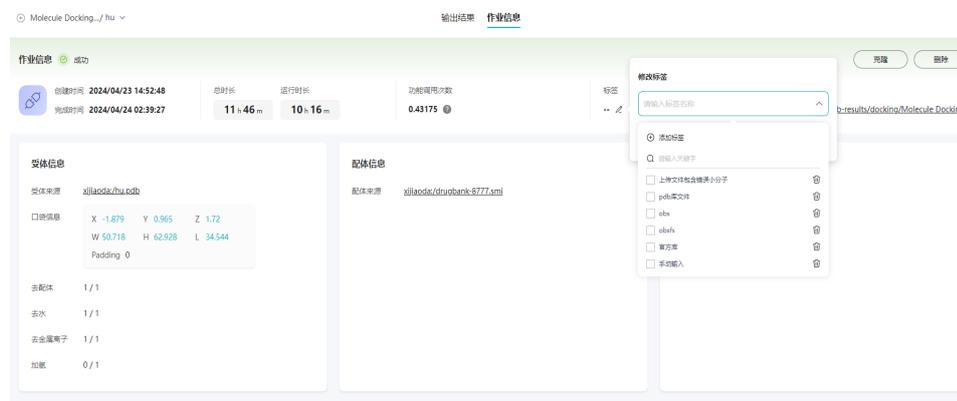
## 作业信息

在作业中心页面，您可以单击具体的作业名称，进入作业信息页面，查看该作业的相关信息，可以克隆或者删除。如果执行失败的作业，可以在该页面查看失败信息。

图 1-162 作业信息页面



作业信息页面也支持标签修改，点击标签 进行修改。



作业执行状态如下所示。

- 等待中：作业创建成功，等待处理中。等待中的作业支持取消。
- 运行中：作业创建成功，执行中。可单击作业名称，进入详情页查看运行详情。“运行中”的作业不允许取消。
- 成功：作业运行成功。可单击作业名称，进入详情页查看运行详情。
- 失败：作业运行失败。可单击作业名称，进入详情页查看运行失败原因。
- 取消中：对“等待中”的作业，执行“取消”操作，取消运行。
- 已取消：取消运行成功。

## 1.12 收藏夹

收藏夹用于收藏用户的任务结果，通过单击相应的作业名称跳转到作业详情页面，可以收藏大小分子，并在收藏夹中查看。收藏夹内分为小分子和大分子，小分子通常为低分子量的化合物，如蛋白质配体；大分子多为蛋白质靶点。

- 小分子涉及分子对接、分子优化、靶点口袋分子设计、CPI、分子生成、分子搜索、FEP、聚类详情页。
- 大分子涉及靶点口袋发现、靶点优化。

## 使用限制

- 下载操作将会产生流量费用，具体可参考[计费说明](#)。
- 如果收藏分子的源作业被删除，收藏列表依旧可以正常展示。
- 收藏夹收藏最多500条。

## 收藏分子

分子对接、分子优化、靶点口袋分子设计、分子生成、靶点口袋发现、靶点优化等模块均可以使用收藏夹功能对结果进行收藏，以下以分子对接模块为例，其他模块的收藏操作可在相应模块的介绍中查阅，请参考[功能模块](#)。

1. 登录盘古辅助制药平台。
2. 单击“进入平台”，进入药物平台操作页面。

- 单击“功能模块 > 苗头化合物发现 > 分子对接模块”，进入分子对接配置页面。根据[分子对接](#)章节完成参数配置。
- 分子对接任务完成后，单击“作业中心 > 作业名称”，进入作业在输出结果页面查看。

图 1-163 查看分子对接结果



- 在输出结果右侧的配体展示列表中，可以单击需要收藏的配体卡片右上方的 进行收藏（在3D视图下），收藏配体列表前两位的配体。单击页面右上角的“只看收藏”，页面筛选出收藏的配体。如果查看全部配体，单击页面右上角的“查看全部”，出现全部配体。

图 1-164 收藏配体



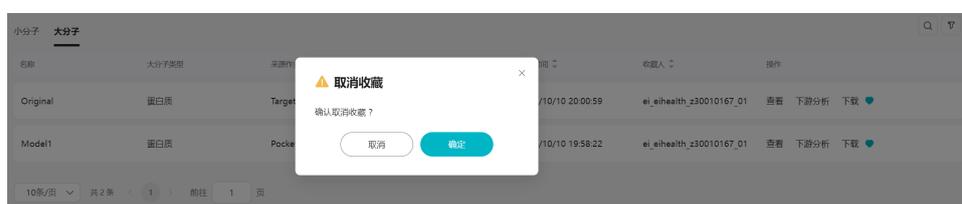
- 在收藏夹查看分子，单击“收藏夹 > 小分子”，查看已收藏的结果。
  - 单击“查看”按钮，可以查看相应的结果信息
  - 单击“下游分析”按钮，可以选择对相应结果进行下游任务的配置。
  - 单击“下载”按钮可以下载相应结果。

图 1-165 收藏结果展示

序号	分子图	名称	SMILES	靶点名称	来源作业	作业类型	收藏时间	收藏人	操作
1		Chloroquine	CCN(CC)CCCC(C)Nc1ccc...	e_3drefine	Molecule Dockin...	分子对接	2024/10/14 15:15...	ei_eihealth_z3001...	查看 下游分析 下载
2		Favipiravir	NC(=O)c1nc(F)cnc1O	e_3drefine	Molecule Dockin...	分子对接	2024/10/14 15:15...	ei_eihealth_z3001...	查看 下游分析 下载

- 当需要取消分子收藏时，单击已收藏的结果右侧的 按钮，弹出取消收藏窗口，单击“确认”，取消对该分子的收藏。

图 1-166 取消分子收藏



## 1.13 相关参数

### 物化性质 ( Physicochemical Property )

#### Molecular Weight (MW):

分子量大小，包含氢原子，最佳分子量在100~600之间。

#### Molecular Refractivity (MR):

总极化率，取决于分子的温度、折射率和压力。

#### Volume:

范德华体积。

#### Density:

密度， $Density = MW / Volume$ 。

#### pKa:

酸解离常数，该值与化合物的结构直接相关。

#### Check Acid:

酸碱性，acid为分子呈酸性，base为碱性。

#### nHA:

Number of hydrogen bond acceptors，氢键受体的数量，最佳在0~12之间。

#### nHD:

Number of hydrogen bond donors，氢键供体的数量，最佳在0~7之间。

**nRot:**

Number of rotatable bonds, 旋转键的数量, 最佳在0~11之间。

**nRing:**

Number of rings, 分子环的数量, 最佳在0~6之间。

**MaxRing:**

Number of atoms in the biggest ring, 最大环的原子数, 最佳在0~18之间。

**nHet:**

Number of heteroatoms, 杂原子的数量, 即非碳原子和氢原子的原子数, 最佳在1~15之间。

**fChar:**

Formal charge, 形式电荷, 最佳在-4~4之间。

**nRig:**

Number of rigid bonds, 刚性键的数量, 最佳在0~30之间。

**Flexibility:**

分子的柔性,  $\text{Flexibility} = \text{nRot} / \text{nRig}$ 。

**stereocenters:**

Number of stereocenters, 立体异构体的数量, 最佳在 $\leq 2$ 。

**TPSA:**

Topological polar surface area, 拓扑极性表面积, 最佳在0~140之间。

**logS:**

The logarithm of aqueous solubility value, 水溶性值的对数, 单位为log mol/L, 最佳区间为-4~0.5 log mol/L。

**logP:**

The logarithm of the n-octanol/water distribution coefficient, 脂溶性值的对数, 单位为log mol/L, 最佳区间在0~3 log mol/L。

## 药物化学 ( Medicinal Chemistry )

**QED:**

- quantitative estimate of drug-likeness, 成药性。
- 结果解释: 大于0.67为易获取, 0.49 ~ 0.67不易获取, 小于0.34为化合物太复杂。

**SAscore:**

- Synthetic accessibility score, 合成可及性分数, 旨在评估分子的合成难易程度。
- 结果解释: 大于等于6为较难合成, 小于6为较易合成。

**SCscore:**

Synthetic complexity score, 合成复杂度分数。

**Fsp<sup>3</sup>:**

- 碳饱和度, 旨在评估分子的空间结构复杂度。
- 结果解释: 大于等于0.42为最佳。

**NPscore:**

- The Natural Product-likeness score, 天然产物的相似度评分。
- 结果解释: 最佳在-5~5之间, 分数越高代表越接近天然产物。

**Lipinski Rule:**

- Lipinski规则,  $MW \leq 500$ ;  $\log P \leq 5$ ;  $nHA \leq 10$ ;  $nHD \leq 5$ 。
- 结果解释: 超过2个及以上不满足以上规则, 代表吸收和渗透性较差。

**Pfizer Rule:**

- Pfizer规则,  $\log P \leq 3$ ;  $TPSA \leq 75$ 。
- 结果解释: 两个都不满足的话, 代表有毒性。

**GSK Rule:**

- GSK规则,  $MW \leq 400$ ;  $\log P \leq 4$ 。
- 结果解释: 满足以上条件, 代表具有较好的成药性特征。

**Golden Triangle:**

- Golden Triangle规则,  $200 \leq MW \leq 500$ ;  $-2 \leq \log D \leq 5$ 。
- 结果解释: 满足以上条件, 代表具有较好的成药性特征。

**PAINS:**

- 频繁命中化合物过滤, PAINS数据库具有480个子结构。
- 结果解释: 数值代表有多少个子结构匹配该数据库, 可以通过DETAIL查看所匹配的子结构特征。

**ALARM NMR Rule:**

- 频繁命中化合物过滤, ALARM NMR数据库具有75个子结构。
- 结果解释: 同PAINS。

**BMS Rule:**

- 频繁命中化合物过滤, BMS数据库具有176个子结构。
- 结果解释: 同PAINS。

**Chelator Rule:**

- 频繁命中化合物过滤, Chelator数据库具有55个子结构。
- 结果解释: 同PAINS。

## 吸收 ( Absorption )

**Caco-2 Permeability:**

- 人体结肠腺癌细胞系 ( Caco-2 ) 用来估计药物渗透性。单位为Log cm/s。
- 结果解释：大于-5.15 log cm/s代表具有较好的渗透性。

#### MDCK Permeability:

- MDCK细胞的表观渗透系数Papp用来估计药物进入人体的吸收效率。单位为cm/s。
- 结果解释：大于 $20 \times 10^{-6}$ cm/s代表具有较好的渗透性， $2 \sim 20 \times 10^{-6}$ cm/s具有一般渗透性，小于 $2 \times 10^{-6}$ cm/s具有较差的渗透性。

#### Pgp-inhibitor:

- Pgp-inhibitor, Pgp抑制。0为非Pgp抑制，1为Pgp抑制。
- 结果解释：预测值的范围在0~1之间。

#### Pgp-substrate:

- Pgp-substrate, Pgp基质。0为非Pgp基质，1为Pgp基质。
- 结果解释：预测值的范围在0~1之间。

#### HIA:

- Human intestinal absorption, 人体肠道吸收指标。HIA大于30%代表具有较好的吸收性 ( HIA+ )，小于30%为 ( HIA- )。
- 结果解释：预测值的范围在0~1之间。

#### F<sub>20%</sub>:

- The human oral bioavailability 20%, 人体口服生物利用度。大于20%为 ( F<sub>20%</sub>+ )，小于20%为 ( F<sub>20%</sub>- )。
- 结果解释：预测值的范围在0~1之间。

#### F<sub>30%</sub>:

- The human oral bioavailability 30%, 人体口服生物利用度。大于30%为 ( F<sub>30%</sub>+ )，小于30%为 ( F<sub>30%</sub>- )。
- 结果解释：预测值的范围在0~1之间。

## 分布 ( Distribution )

#### PPB:

- Plasma protein binding, 血浆蛋白结合度。结合度的多少会影响药物的渗透情况。
- 结果解释：结合度小于90%比较合适。

#### VD:

- Volume Distribution, 体积分布，单位为L/kg。
- 结果解释：VD值在0.04~20L/kg之间较好。

#### BBB Penetration:

- blood-brain barrier Penetration, 血脑屏障。作用于中枢神经系统的药物需要穿过血脑屏障 (BBB) 才能达到其分子靶点。相反，对于具有外围靶点的药物，不需

要BBB渗透以避免中枢神经系统副作用，单位为cm/s。logBB 大于-1为 ( BBB+ )，logBB小于等于-1为 ( BBB- )，数值在0~1之间。

- 结果解释：预测值的范围在0~1之间。

#### Fu:

- The fraction unbound in plasms，未结合血清蛋白的比例。药物与血清蛋白的结合程度越高，渗透性越差。
- 结果解释：未结合比例大于等于5%效果较好。

## 排泄 ( Excretion )

#### CL:

- The clearance of a drug，药物的清除率。单位为ml/min/kg。
- 结果解释：清除率大于等于5 ml/min/kg较好。

#### T<sub>1/2</sub>:

- 药物的半衰期。T<sub>1/2</sub>大于3为 ( T<sub>1/2</sub>+ )，小于等于3为 ( T<sub>1/2</sub>- )。
- 结果解释：数值范围在0~1之间，数值越高，代表半衰期越长，越不容易清除。

## Vina score

靶点口袋分子设计后的分子的vina score分数。

## similarity

靶点口袋分子设计后的分子与原始分子的相似度，计算的是ECFP4 Tanimoto相似度。

## score

优化后小分子的综合打分。

## Vina Score1

分子优化如果添加了靶点，将会计算对接结合能，并按照Vina Score进行排序，Vina Score Top1分数。

## Vina Score2

分子优化如果添加了靶点，将会计算对接结合能，并按照Vina Score进行排序，Vina Score Top2分数。

## Vina Score3

分子优化如果添加了靶点，将会计算对接结合能，并按照Vina Score进行排序，Vina Score Top3分数。

## 代谢 ( Metabolism )

#### Metabolism

- 药物代谢反应过程可以分为两大类：第一阶段（氧化反应），第二阶段（共轭反应），人类细胞色素P450家族与第一阶段的氧化反应有关，其中1A2、3A4、2C9同工酶尤为重要，大部分集中在肝脏。0为非抑制剂/非基质，1为抑制剂/基质，数值范围在0~1之间。
- 结果解释：预测值的范围在0~1之间。

#### **CYP1A2 inhibitor**

- 细胞色素P450 1A2亚型抑制剂，0为非抑制剂，1为抑制剂。
- 结果解释：预测值的范围在0~1之间。

#### **CYP2C9 inhibitor**

- 细胞色素P450 2C9亚型抑制剂，0为非抑制剂，1为抑制剂。
- 结果解释：预测值的范围在0~1之间。

#### **CYP3A4 inhibitor**

- 细胞色素P450 3A4亚型抑制剂，0为非抑制剂，1为抑制剂。
- 结果解释：预测值的范围在0~1之间。

#### **CYP2C19 inhibitor**

- 细胞色素P450 2C19亚型抑制剂，0为非抑制剂，1为抑制剂。
- 结果解释：预测值的范围在0~1之间。

#### **CYP2D6 inhibitor**

- 细胞色素P450 2D6亚型抑制剂，0为非抑制剂，1为抑制剂。
- 结果解释：预测值的范围在0~1之间。

## **生物活性（Bioactivity）**

### **β-secretase 1 inhibitor:**

- beta-分泌酶抑制剂。0为非抑制剂，1为抑制剂，数值范围在0~1之间。
- 结果解释：预测值的范围在0~1之间。

### **HIV inhibitor:**

- HIV 复制抑制剂。0为非抑制剂，1为抑制剂，数值范围在0~1之间。
- 结果解释：预测值的范围在0~1之间。

## **毒性（Toxicity）**

### **hERG Blockers:**

- The human ether-a-go-go related gene, hERG阻断心悸、昏厥甚至猝死。IC50>10μm或者在10μm小于50%抑制率定义为hERG-（0），IC50<10μm或者在10μm时大于50%抑制率定义为hERG+（1）。
- 0为hERG阴性，1为hERG阳性，数值范围在0~1之间。
- 结果解释：预测值的范围在0~1之间。

### **H-HT:**

- The human hepatotoxicity, 药物肝损伤。0为H-HT阴性, 1为H-HT阳性, 数值范围在0~1之间。

结果解释: 预测值的范围在0~1之间。

#### **DILI:**

- Drug-induced liver injury (DILI), 药物肝损伤。0为DILI阴性, 1为DILI阳性, 数值范围在0~1之间。

- 结果解释: 预测值的范围在0~1之间。

#### **AMES Toxicity:**

- The Ames test for mutagenicity, 致突变试验。0为AMES阴性, 1为AMES阳性, 数值范围在0~1之间。

- 结果解释: 预测值的范围在0~1之间。

#### **FDAMDD:**

- 药物毒性计量阈值。单位为mmol/kg-bw/day, 小于等于0.011 mmol/kg-bw/day 为FDAMDD阳性, 大于0.011 mmol/kg-bw/day 为FDAMDD阴性。

- 结果解释: 预测值的范围在0~1之间。

#### **Skin Sensitization:**

- 皮肤过敏性。药物皮肤过敏可诱发过敏性接触性皮炎, 是一种重要的指标。0非不皮肤过敏, 1为皮肤过敏, 数值在0~1之间。

- 结果解释: 预测值的范围在0~1之间。

#### **Carcinogenicity:**

- 致癌性。0为非致癌物, 1为致癌, 数值在0~1之间。

- 结果解释: 预测值的范围在0~1之间。

#### **Eye Corrosion/Irritation:**

- 眼睛腐蚀性/刺激性。角膜和结膜组织构成眼睛的前表面, 因此角膜结膜组织直接暴露在空气中, 容易受到化学物质的伤害。0为非眼睛腐蚀或者刺激性, 1为眼睛腐蚀或者刺激性, 数值在0~1之间。

- 结果解释: 预测值的范围在0~1之间。

## **环境毒性 ( Environmental Toxicity )**

#### **Bioconcentration Factors:**

生物浓缩系数 ( BFC )。生物体内某种化合物的浓度同水中该物质的浓度比值, 单位为 $\log_{10}(\text{L}/\text{kg})$ 。

#### **IGC<sub>50</sub>:**

48 hour Tetrahymena pyriformis IGC50 ( 50% Inhibition Growth Concentration ) , 导致50% Tetrahymena pyriformis生长抑制, 在48小时后的化学物质浓度。单位为 $-\log_{10}[(\text{mg}/\text{L})/(1000 \cdot \text{MW})]$ 。

#### **LC<sub>50</sub>FM:**

96 hour fathead minnow LC50, 导致50% fathead minnow死亡, 在96小时后的化学物质浓度。单位为 $-\log_{10}[(\text{mg}/\text{L})/(1000 \cdot \text{MW})]$ 。

### LC<sub>50</sub>DM:

48 hour Daphnia magna LC50, 导致50% Daphnia magna死亡, 在48小时后的化学物质浓度。单位为 $-\log_{10}[(\text{mg/L})/(1000 \cdot \text{MW})]$ 。

## Tox21 通路 (Tox21 Pathway)

### NR-AR:

- Androgen receptor, 雄激素受体。在AR依赖性前列腺癌和其他相关雄激素相关疾病中其关键作用, 内分泌干扰物与AR的相互作用, 会破坏正常的内分泌功能, 因此需要了解化学物质对ER信号通路的影响。0为AR无活性, 1为AR有活性。数值在0~1之间。
- 结果解释: 预测值的范围在0~1之间。

### NR-AR-LBD:

- 雄激素受体配体结合区。0为AR-LBD无活性, 1为AR-LBD有活性。数值在0~1之间。
- 结果解释: 预测值的范围在0~1之间。

### NR-AhR:

- The Aryl hydrocarbon Receptor, 芳香烃受体。AhR是一种配体激活性转录因子。当与多环芳烃、卤代芳烃等配体结合后, 可调控一系列基因的表达。0为AhR无活性, 1为AhR有活性。数值在0~1之间。
- 结果解释: 预测值的范围在0~1之间。

### NR-Aromatase:

- Aromatase, 芳香化酶, 是雌激素生物合成的关键酶。0为Aromatase无活性, 1为Aromatase有活性。数值在0~1之间。
- 结果解释: 预测值的范围在0~1之间。

### NR-ER:

- Estrogen receptor, 雌激素受体。在发育和新陈代谢中起重要作用, 内分泌干扰物与ER的相互作用, 会破坏正常的内分泌功能, 因此需要了解化学物质对ER信号通路的影响, 0为Estrogen receptor无活性, 1为Estrogen receptor有活性。数值在0~1之间。
- 结果解释: 预测值的范围在0~1之间。

### NR-ER-LBD:

- 雌激素受体配体结合区。0为ER-LBD无活性, 1为ER-LBD有活性。数值在0~1之间。
- 结果解释: 预测值的范围在0~1之间。

### NR-PPAR-gamma:

- peroxisome proliferator-activated receptors PPAR, 过氧化物酶体增殖物激活受体。PPAR是一类由配体激活的核转录因子, PPAR超家族有3个亚型:即PPAR- $\alpha$ 、PPAR- $\delta$  (又称PPAR- $\beta$ 或NUC-1) 和PPAR- $\gamma$ , 分别由不同的基因编码, PPAR- $\gamma$ 受体 (格列酮受体)。参与调节糖脂代谢。0为PPAR-gamma无活性, 1为PPAR-gamma有活性。数值在0~1之间。

- 结果解释：预测值的范围在0~1之间。

#### SR-ARE:

- antioxidant response element ARE，抗氧化反应元件。ARE与Nrf2结合，启动下游大量的抗氧化酶基因的转录。0为ARE无活性，1为ARE有活性。数值在0~1之间。
- 结果解释：预测值的范围在0~1之间。

#### SR-ATAD5:

- ATPase family AAA domain-containing protein 5，三磷酸腺苷酶家族蛋白5。ATAD5会随着各种类型的DNA损伤而在蛋白质水平上增长。0为ATAD5无活性，1为ATAD5有活性，数值在0~1之间。
- 结果解释：预测值的范围在0~1之间。

#### SR-HSE:

- Heat shock factor，热激蛋白。HSE是一类功能性相关蛋白质，当细胞受到升高温度或其他胁迫时它们的表达就会增长。0为HSE无活性，1为HSE有活性，数值在0~1之间。
- 结果解释：预测值的范围在0~1之间。

#### SR-MMP:

- Mitochondrial membrane potential，线粒体膜电位。线粒体是动植物细胞生成ATP的主要地点，是促进细胞能量转换、参与细胞凋亡的重要细胞器。线粒体在产生能量时会把电势差储存于线粒体内膜，在内膜两侧，若质子及其他离子浓度的不对称分布就会形成线粒体膜电位，即MMP，测量活细胞中的MMP通常用于评估化学物质对线粒体功能的影响。0为MMP无活性，1为MMP有活性，数值在0~1之间。
- 结果解释：预测值的范围在0~1之间。

#### SR-p53:

- p53，一种肿瘤抑制因子，控制细胞周期的启动。在所有恶性肿瘤中，50%以上会出现该基因的突变。0为p53无活性，1为p53有活性，数值在0~1之间。
- 结果解释：预测值的范围在0~1之间。

## 毒效团（Toxicophore Rules）

#### Acute Toxicity Rule:

- 有20个子结构，含有该子结构可能在口服给药期间引起急性毒性。
- 结果解释：数值代表有多少个子结构匹配此数据库，可以通过DETAIL查看所匹配的子结构特征。

#### Genotoxic Carcinogenicity Rule:

- 117个子结构，含有该子结构可能通过遗传毒性引起致癌性或者致突变性。
- 结果解释：数值代表有多少个子结构匹配此数据库，可以通过DETAIL查看所匹配的子结构特征。

#### NonGenotoxic Carcinogenicity Rule:

- 23个子结构，含有该子结构可能通过非遗传毒性引起致癌性或者致突变性。  
结果解释：数值代表有多少个子结构匹配此数据库，可以通过DETAIL查看所匹配的子结构特征。

**Skin Sensitization Rule:**

- 155个子结构，含有该子结构可能会导致皮肤损伤。  
结果解释：数值代表有多少个子结构匹配此数据库，可以通过DETAIL查看所匹配的子结构特征。

**Aquatic Toxicity Rule:**

- 99个子结构，含有该子结构可能会对液体（包括水）造成毒性。
- 结果解释：数值代表有多少个子结构匹配此数据库，可以通过DETAIL查看所匹配的子结构特征。

**NonBiodegradable Rule:**

- 19个子结构，含有该子结构可能会不可生物降解。
- 结果解释：数值代表有多少个子结构匹配此数据库，可以通过DETAIL查看所匹配的子结构特征。

**SureChEMBL Rule:**

- 164个子结构，含有该子结构可能会具有不好的药物化学结构。  
结果解释：数值代表有多少个子结构匹配此数据库，可以通过DETAIL查看所匹配的子结构特征。

**FAF-Drugs4 Rule:**

- 154个子结构，含有该子结构可能会有毒性。
- 结果解释：数值代表有多少个子结构匹配此数据库，可以通过DETAIL查看所匹配的子结构特征。

# 2 用户指南（基因平台）

## 2.1 欢迎使用医疗智能体服务

医疗智能体（EIHealth）平台是基于华为云AI和大数据技术优势，为基因组分析、药物研发和临床研究三个领域提供的专业AI研发平台。

平台为您提供高性能、高可靠性、高性价比的计算、存储、分析和AI能力支持，让科研过程标准化、可执行，提升医疗健康的服务能力和普惠水平。同时通过丰富的加速技术，实现算法工具的数倍加速，加速生物医疗大数据的分析。

平台由项目管理、数据管理、作业、工具、开发环境、镜像等核心部件组成，各个部件沉淀了丰富的技术细节和人性化的设计。

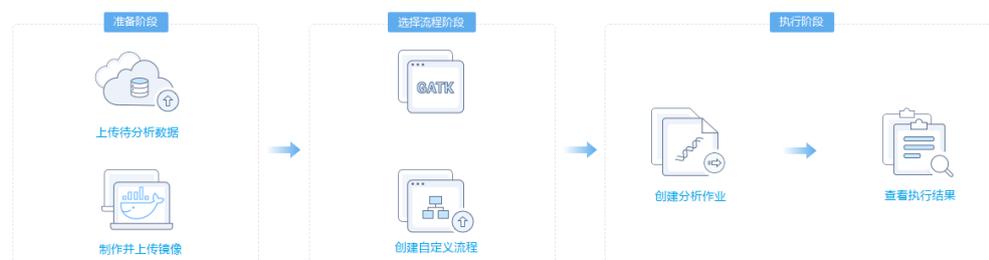
初学者能够基于页面可视化的完成数据管理、复现业内的分析流程和算法。资深从业者能够基于镜像打造自己的分析流程。

### 使用流程

EIHealth提供了个性化分析流程的搭建和管理操作。您可以将生物信息学软件封装为应用，并将其编排调度，形成自定义分析流程。

同时集成了基于开源的Jupyter Notebook，可为您提供在线的开发和调试工具，用于编写和调测模型训练代码。

图 2-1 使用流程



## 服务声明

[医疗智能体服务声明](#)，将为您介绍在使用华为云医疗智能体服务时所享有的权利、履行的义务和责任。

## 2.2 关键概念

### 镜像

运行生物信息学软件，往往由于不同的操作系统（Windows、Linux、Mac等）原因，无法实现统一的运维管理。同时，这些软件具有不同的版本和软件包，安装、使用过程复杂。将生物信息学软件封装成Docker镜像，可以使程序在不同的环境中运行，并通过EIHealth平台的镜像管理，实现高效的调用，极大方便了软件的安装和运行。

Docker镜像是一个模板，是容器应用打包的标准格式，在部署容器化应用时可以指定镜像。例如一个Docker镜像可以包含一个完整的Ubuntu操作系统环境，里面仅安装了用户需要的应用程序及其依赖文件。EIHealth平台使用容器镜像服务（Software Repository for Container，简称SWR）进行简单易用、安全可靠的镜像管理。

### 应用

EIHealth平台的每一个分析作业都依托于应用运行。应用是生物信息学软件和运行该软件所依赖的运行环境的镜像封装，应用可以独立使用，也可以将多个应用编排入流程串联使用。

您可以在项目的应用列表中，查看隶属于该项目的应用，也可以通过搜索应用名称快速查找所需应用。应用列表展示了应用的名称、版本、创建者、修改时间、创建时间和可执行的操作。

详细的应用创建和使用请参见[工具管理](#)。

- 创建应用  
应用是生物信息学软件的镜像封装，您可以制作软件镜像并上传至平台，并通过“[新建应用](#)”引入相关软件。
- 导入应用  
应用按项目进行划分，隶属于不同项目的应用，可以通过“[导入应用](#)”的方法，导入至自己的项目中使用。

### 流程

EIHealth中的分析流程包含分析过程中所需应用的执行信息和数据的输入、输出等参数定义。

分析流程至少由一个应用组成，在多个应用构成的流程中，一个应用的输出作为另一个应用的输入，流程中的各个应用由其前后顺序形成完整的计算 workflow。

分析流程通过流程设计器创建，创建好的流程将存储于项目中。同时，您也可以通过“[导入流程](#)”的方法，将隶属于其他项目的流程导入至自己的项目中。

创建好的流程显示在流程列表中，您可以基于这些流程创建分析作业。

详细的流程创建和运行请参见[新建流程](#)。

## 2.3 准备工作

### 2.3.1 购买服务

#### 前提条件

- **注册华为账号并开通华为云**，并进行**账号实名认证**。购买服务时，账号不能处于欠费或冻结状态。
- 购买服务时，需使用华为主账号购买。如果不确定账号类型，可登录**账号中心**查看所属账号是主账号还是子账号（IAM账号）。  
在账号中心左侧导航栏中有“我的主账号”页签，则说明是子账号；没有“我的主账号”页签，则说明是主账号。

图 2-2 查看账号信息



#### 购买服务

医疗智能体平台提供包年包月、按需两种计费方式，您可以按需求进行选择。

关于计费模式的详细介绍，请参见**计费说明**。

1. 登录华为云**管理控制台**，在左侧服务列表中选择“人工智能 > 医疗智能体 EIHealth”。

图 2-3 服务列表



2. 进入医疗智能体“总览”页面，单击“EIHealth”。

图 2-4 购买 EIHealth 平台



3. 根据实际需求选择计费方式。

图 2-5 选择计费方式



- 区域：请就近选择靠近您业务的区域，可减少网络延迟，提高访问速度。不同区域云服务产品之间内网互不相通。
- 平台规格：选择“基因平台”或“药物平台”。
- 平台计费方式：选择“包年包月”或“按需”计费。
- 购买时长：如果选择的“包年包月”计费，根据实际需求选择购买时长；如果选择的“按需”计费，无需选择购买时长。  
勾选自动续费后，系统将在产品到期前自动续费，无需用户再手动操作。按月购买，自动续费周期为一个月。

4. 单击“下一步”，确认订单详情。

5. 单击“立即购买”，并支付订单，完成服务购买。

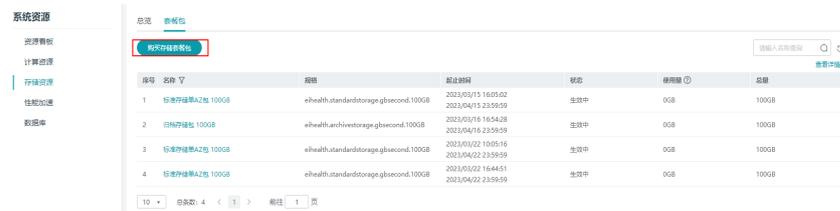
购买成功后，需要大约30分钟时间完成服务自动部署，请耐心等待。部署完成后，您可以使用注册的华为账号使用医疗智能体平台。

## 2.3.2 购买存储套餐包

存储资源支持用户购买套餐包。存储资源包不能退订，不能抵扣已经产生的用量。您可以单击右上角“查询详情”，查看套餐包的具体使用量。

1. 在平台右上角单击用户名，选择“系统资源” > “存储资源”。
2. 在存储资源页面选择“套餐包”，单击“购买存储套餐包”。

图 2-6 购买存储套餐包

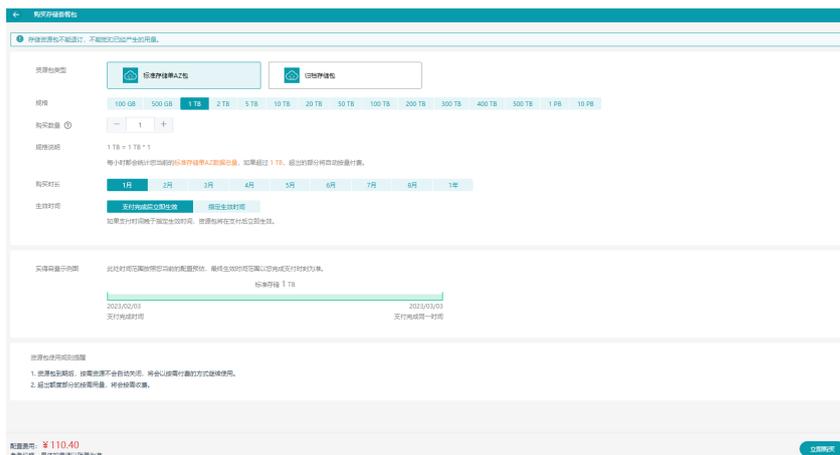


3. 填写套餐包的信息。

表 2-1 参数说明

参数	说明
资源包类型	支持标准存储单AZ包、归档存储包。 标准存储单AZ包适合数据分享、内容分享、热点对象应用场景。归档存储适合档案数据、医疗影像场景。详细可参考 <a href="#">存储类别</a> 。
规格	套餐包支持的购买规格。
购买数量	购买套餐包的数量。取值范围：1-50。
购买时长	购买套餐包的时长。最短为一个月，最长为一年。
生效时间	选择购买套餐包后的生效时间。目前支持支付完成后立即生效和指定生效时间两种方式。 如果支付时间晚于指定生效时间，资源包将在支付后立即生效。

图 2-7 购买套餐包



4. 设置完成后，单击“立即购买”，然后单击“确认”。
  5. 在云服务订单页面，单击“确认付款”。
- 购买成功后，可以在“套餐包”页面查看套餐包信息。

### 说明

- 资源包到期后，按需资源不会自动关闭，将会以按需付费的方式继续使用。
- 超出套餐包部分的用量，将会按需收费。
- 若平台先到期，套餐包还没到期，则不退套餐包剩余的费用。如果平台续费成功后，套餐包可以继续使用。
- 资源包不能抵扣已经产生的用量。

## 2.3.3 购买系统资源

医疗智能体平台购买完成后，单击“进入平台”，在平台右上角单击用户名，选择“系统资源”。

您可以根据实际需求选择购买计算资源、性能加速、数据库。在后续使用过程中，也可根据使用情况随时购买资源。

## 资源看板

在“资源看板”中，您可以实时监控计算资源、存储资源、性能加速、数据库的使用情况。

您可以选择统计“近1小时”、“近3小时”、“近12小时”、“近24小时”、“近7天”、“近30天”的资源使用情况，在“周期”中设置统计周期时长，在“方法”中选择“最大值”、“平均值”、“最小值”。如图中，上箭头是最大值，下箭头是最小值，虚线右边的数据是平均值。

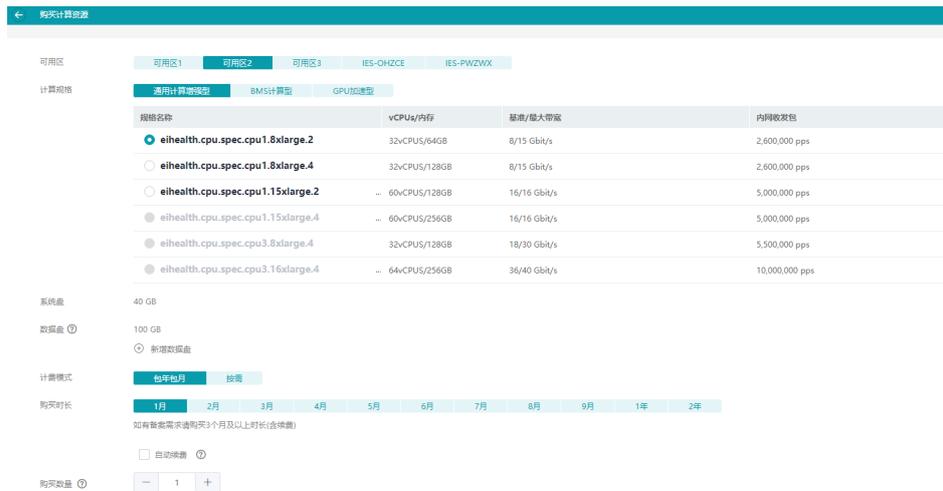
图 2-8 资源看板



## 购买计算资源

1. 在“计算资源”页面，单击“购买计算资源”。
2. 选择“可用区”、“计算规格”、“系统盘”、“数据盘”、“计费模式”、“购买时长”、“购买数量”。

图 2-9 购买计算资源



- 可用区：选择可用区。
- 计算规格：分为通用计算增强型、内存优化型、BMS计算型、磁盘增强型、超大内存型、GPU加速型，根据需求选择类型，并勾选规格名称。
- 系统盘：40GB，不可修改。
- 数据盘：数据盘可用于在创建作业时，开启“本地盘加速”，默认100GB，可按使用需要进行扩增。
- 计费模式：选择“包年包月”或“按需”计费。
- 购买时长：如果选择的“包年包月”计费，根据实际需求选择购买时长；如果选择的“按需”计费，无需选择购买时长。  
开通自动续费后，系统将在产品到期前自动续费，续费周期为一个月，无需用户再手动操作。
- 购买数量：设置购买数量，最多10个。

3. 单击“立即购买”。
4. 在确认购买弹窗中单击“确认”。

计算资源购买后，在操作列单击“更多”，可以执行开机、关机、重启、标签管理、调度设置操作。

图 2-10 计算资源列表



## 计算资源标签管理

**标签管理：**可通过设置系统内置标签“copy-in”，使作业在该计算节点上运行。用户也可以自定义标签，创建应用时，通过设置标签指定应用运行在具有相同标签的计算节点上，实现运行作业按标签进行调度。

**调度设置：** 设置为“不可调度”后，作业将无法运行到此计算节点。

图 2-11 标签管理



## 计算资源节点自动扩缩容

为了满足客户业务对节点的需求，医疗智能体（EIHealth）平台支持用户设置节点自动扩缩容策略。

### 创建扩容策略

1. 在平台右上角用户名中选择“系统资源>计算资源>节点伸缩策略”，单击“创建扩容策略”，或者在“节点”页面选择某个节点，单击操作列“更多>配置扩容策略”进入创建扩容策略页面。
2. 在创建扩容策略页面配置相关信息。

表 2-2 参数说明

参数	说明
策略名称	设置策略名称。策略名称长度1-64，只支持字母，数字，中划线和下划线。 如果从操作列“更多>配置扩容策略”进入创建页面，策略名称会有默认值，默认名称会携带规格信息。
计费模式	按需计费。扩容的节点只支持按需计费。
可用区	选择可用区。
计算规格	选择计算节点类型及对应的规格。
系统盘	系统盘固定为40GB。
数据盘	默认为100GB，可以根据需要选择添加数据盘。 当系统盘和数据盘的总配额为3000GB时，如需增加，请联系管理员。
节点数上限	设置自动扩容的节点数上限。 取值范围为0-50。
节点数下限	设置自动扩容的节点数下限。 取值范围为0-50。

参数	说明
CPU执行规则	设置CPU自动扩容的条件。 <ul style="list-style-type: none"> <li>• 启用规则：默认开启。</li> <li>• 触发条件：设置CPU分配率达到的条件，当满足设置的条件时进行自动扩容。取值范围：1-100。默认值：80。</li> <li>• 执行动作：设置当达到触发条件时，增加的节点个数。取值范围：1-50。默认值：1。</li> </ul>
内存执行规则	设置内存自动扩容的条件。 <ul style="list-style-type: none"> <li>• 启用规则：默认开启。</li> <li>• 触发条件：设置内存分配率达到的条件，当满足设置的条件时进行自动扩容。取值范围：1-100。默认值：80。</li> <li>• 执行动作：设置当达到触发条件时，增加的节点个数。取值范围：1-50。默认值：1。</li> </ul>

3. 单击“确定”，可以在节点伸缩策略页面查看创建的扩容策略。

扩容策略最多支持10个。当扩容策略创建成功后，您可以根据需求，选择是否开启“自动扩缩容”。如果开启，系统将会根据配置的策略条件，满足条件后，自动扩缩容节点。关闭后不会进行扩缩容。

图 2-12 节点伸缩策略



也可以通过单击某个策略操作列的“编辑”或者“删除”，对已有的策略进行修改或者删除操作。如果删除，则会同步删除节点。如果编辑，不支持修改可用区、节点规格、系统盘和数据盘，其他参数均可修改。

单击某个策略操作列的“事件”，查看该策略的事件信息。事件保存时间为30天，30天后自动清除数据，查询时间最多15天。

单击策略名称前面的箭头，可以查看策略的关联节点列表、执行规则和伸缩历史记录信息。

图 2-13 查看节点执行规则



图 2-14 查看节点伸缩历史



图 2-15 查看关联节点列表



### 配置缩容策略

缩容策略只支持统一配置，且不会对手动购买的计算资源节点缩容。

1. 在平台右上角用户名中选择“系统资源>计算资源>节点伸缩策略”。
2. 单击“配置缩容策略”。

图 2-16 配置缩容策略

缩容策略
✕

! 不会对手动购买的节点进行缩容，缩容策略只支持统一配置。

\* 空置时间  分钟  
节点连续空闲多长时间进行缩容，默认10分钟

\* 缩容阈值  %  
当节点资源使用量低于多少时认为是空闲（CPU与内存同时满足）

高级参数 ^

\* 冷却时间  分钟  
扩容执行后多久能再次判断是否缩容

分钟  
节点删除后多久能再次判断是否缩容

分钟  
缩容失败后多久能再次判断是否缩容

\* 缩容并发数  个  
最多支持多少个空闲节点同时缩容

\* 检查间隔  分钟  
节点被判定为不可缩容后再次检查的时间间隔

确定
取消

表 2-3 缩容策略参数说明

参数	说明
空置时间	节点连续空闲多长时间进行缩容，默认10分钟。取值范围：1-10080。
缩容阈值	当节点资源使用量低于多少时认为是空闲（CPU与内存同时满足）。取值范围：1-100。
冷却时间	冷却时间需设置三部分。 <ul style="list-style-type: none"> <li>• 扩容执行后多久能再次判断是否缩容。取值范围：5-10080。</li> <li>• 节点删除后多久能再次判断是否缩容。取值范围：1-10080。</li> <li>• 缩容失败后多久能再次判断是否缩容。取值范围：1-10080。</li> </ul>
缩容并发数	最多支持多少个空闲节点同时缩容。取值范围：1-50。
检查间隔	节点被判定为不可缩容后再次检查的时间间隔。取值范围：1-10080。

3. 单击“确定”。

如果有作业在运行的节点，不会被扩容，不影响客户业务。

### 说明

对于关机的节点，空闲时间是单独设置的，默认为20分钟。如果超过20分钟，系统将直接扩容，不受扩容策略影响。

## 存储资源

在“存储资源>总览”页面，显示存储资源的“序号”、“名称”、“规格”、“订购时间”、“使用量”。

如果存储资源被冻结，则不能使用医疗智能体（EIHealth）平台。

图 2-17 存储资源

序号	名称	规格	订购时间	使用量
1	医疗智能体多人存储	eihealth.storage.spec.multiaz	2022/07/19 11:54:54	16.52TB
2	医疗智能体归档存储	eihealth.storage.spec.archive	2022/12/01 23:09:14	450.15GB

## 购买性能加速

对于涉及频繁读写场景的任务建议购买性能加速，用于创建分析作业时开启IO加速。

1. 在“性能加速”页面，单击“购买性能加速”。
2. 选择“存储空间大小”、“计费模式”、“购买时长”、“购买数量”。

图 2-18 购买性能加速

存储空间大小: 100 TB

计费模式: 包年包月 | 按需

购买时长: 1月 | 2月 | 3月 | 4月 | 5月 | 6月 | 7月 | 8月 | 9月 | 1年 | 2年

自动续费

购买数量: 1

- 存储空间大小：设置存储空间数值。
- 计费模式：选择“包年包月”或“按需”计费。
- 购买时长：如果选择的“包年包月”计费，根据实际需求选择购买时长；如果选择的“按需”计费，无需选择购买时长。  
勾选自动续费后，系统将在产品到期前自动续费，续费周期为一个月，无需用户再手动操作。
- 购买数量：设置购买数量。

3. 单击“立即购买”。

4. 在确认购买弹窗中单击“确认”。

性能加速资源购买后，可以设置性能加速作业配额。在操作列单击“配置”，设置每个性能加速实例上允许运行的最大作业个数，默认为8个/10T。

在业务量较低情况下，可以支持客户删除一些性能加速资源。但在可调度的情况下删除有可能会对客户业务中断，所以需要先设置不可调度，再删除。性能加速资源购买后，可以设置性能加速调度设置。在操作列单击“更多>调度设置”，设置每个性能加速实例的调度状态。

**说明**

- 如果购买的性能加速是包年包月计费，则支持续订，退订以及释放。
- 如果购买的性能加速是按需计费，则支持单击操作列“删除”，进行资源删除。
- 性能加速节点有作业运行时，无法删除、退订。
- 当作业中部分task无加速，部分io加速时，当性能加速点设置为不可调度时，整个作业处于等待中状态。

图 2-19 性能加速节点信息

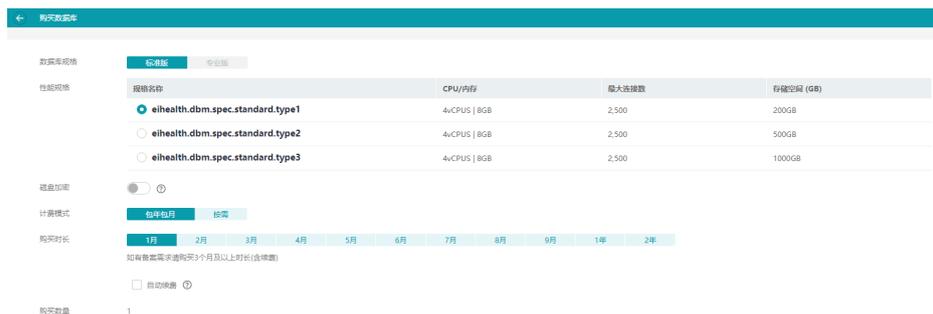


## 购买数据库

使用数据库功能前，需要先购买数据库，数据库只能购买一个。

1. 在“数据库”页面，单击“购买数据库”。
2. 选择“数据库规格”、“性能规格”、“磁盘加密”、“计费模式”、“购买时长”、“购买数量”。

图 2-20 购买数据库



- 数据库规格：选择“标准版”。
- 性能规格：根据您的需求选择规格。
- 磁盘加密：选择加密后会提高数据安全性，但对数据库读写性能有少量影响，请按照您的使用策略进行选择。
- 计费模式：选择“包年包月”或“按需”计费。
- 购买时长：如果选择的“包年包月”计费，根据实际需求选择购买时长；如果选择的“按需”计费，无需选择购买时长。

勾选自动续费后，系统将在产品到期前自动续费，续费周期为一个月，无需用户再手动操作。

- 购买数量：1个，不可修改。
- 3. 单击“立即购买”。
- 4. 在确认购买弹窗中单击“确认”。

## 续订、退订、释放资源

购买资源选择包年包月计费模式，可以进行续订、退订或释放。

- 续订、退订资源  
勾选已购买的资源，在页面上单击“续订”或“退订”。

图 2-21 续订或退订资源

序号	名称	规格	数据盘	订购时间	计费模式	购买时长	运行状态	已用容量	操作
1	EHealth Database Star	ehealth_dbm.spec.star (4vCPU   8GB)	40GB	2021/09/10 20:23...	包年包月	--	正常	4.61GB	续订 更多

- 释放资源  
当购买的资源处于冻结状态时，可以选择释放资源。勾选已购买的资源，在页面上单击“释放”。

图 2-22 释放资源

序号	名称	可用区	规格	订购时间	计费模式	购买时长	已用容量	最大容量	作业数	运行状态	操作
1	sf-turbo-35fe5314-101	AZ3	ehealth.storage1...	2022/08/27 17:46...	包年包月 (已冻结)	--	0.00GB	300GB	15	冻结	续订   配置 更多

## 删除资源

购买资源时选择按需计费模式，如果资源不再使用，可以删除资源。在已购买资源的“操作”列，单击“删除”。

图 2-23 删除资源

序号	名称	规格	数据盘	订购时间	计费模式	购买时长	运行状态	已用容量	操作
1	EHealth Database Star	ehealth_dbm.spec.star (4vCPU   8GB)	40GB	2021/10/28 20:22...	按需	--	正常	5.60GB	删除

## 查找资源

- 普通搜索：在搜索框中输入关键字，单击 按钮查找资源。

图 2-24 查找资源

- 高级搜索：计算资源支持按照名称、运行状态、计费模式、规格名称、标签进行搜索。输入某一项搜索条件后，单击“搜索”即可查找出相关信息。

图 2-25 高级搜索



## 资源统计

计算资源支持选择一部分节点，统计出对应节点的资源使用量。

1. 在计算资源页面，选择需要统计的节点，单击“资源统计”。

图 2-26 选择统计节点



### 说明

- 选择的节点不能跨页。
2. 配置资源统计的时间、周期和方法。

图 2-27 配置资源统计信息



3. 单击“下载”。
- 如选择统计方法为“最大值”，则选择所有取值的最大值；如选择“平均值”，则选择所有取值的平均值。

## 查看计算资源节点监控

您可以在计算资源页面，单击节点“监控”列后面的图标，直接查看资源节点的监控信息。上箭头是最大值，下箭头是最小值，虚线右边的数据是平均值。

图 2-28 查看节点监控信息



## 2.3.4 获取认证信息

### 获取 AK/SK

AK/SK（Access Key ID/Secret Access Key）即访问密钥，包含访问密钥ID（AK）和秘密访问密钥（SK）两部分，华为云通过AK识别用户的身份，通过SK对请求数据进行签名验证，用于确保请求的机密性、完整性和请求者身份的正确性。

您在使用命令行工具时，需要使用AK/SK进行身份验证。

### 操作步骤

1. 登录[华为云管理控制台](#)，鼠标指向页面右上角的用户名，在下拉列表中单击“我的凭证”。

图 2-29 我的凭证入口



2. 在“我的凭证”页面中选择“访问密钥”页签。单击“新增访问密钥”，按操作指引获取认证账号的AK/SK，请妥善保管AK/SK信息。

图 2-30 访问密钥



### 说明

- 每个用户仅允许新增两个访问密钥。
- 为保证访问密钥的安全，访问密钥仅在初次生成时自动下载，后续不可再次通过管理控制台页面获取。请在生成后妥善保管。

## 获取区域名称

区域是一个地理区域的概念，由于带宽原因，会在多个地区建立数据中心，提供服务。

医疗智能体服务部署在华北-北京四，区域名称为cn-north-4；部署在华东-上海一，区域名称为cn-east-3。

## 获取平台 ID

平台ID与“[命令行工具 > 初始化配置](#)”中的platform-id对应。

1. 登录医疗智能体平台。
2. 在右上角用户名中选择“个人设置”，获取平台ID。

图 2-31 平台 ID

#### 系统信息

平台 ID e53ft 

## 2.4 用户管理

### 2.4.1 用户管理介绍

EIHealth平台支持两种类型的用户管理，可以帮助您安全地控制平台的访问和使用权限。

表 2-4 用户管理类型

类型	说明
系统级别用户管理	系统级的角色配置，可创建平台的子用户，并为其分配权限。
项目级别用户管理	资源级的角色配置，以项目为粒度对数据、分析作业、开发环境和镜像进行分组，以使用户通过项目进行资源的访问、共享和协作。详细介绍请参见 <a href="#">项目管理</a> ，项目级用户不同角色对应的权限请参考 <a href="#">成员角色和权限</a> 。

图 2-32 系统级别用户管理

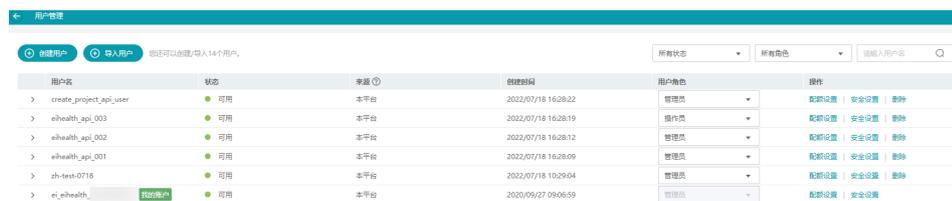


图 2-33 项目级别用户管理



购买平台的账号是资源的归属以及使用计费的主体，对其所拥有的资源具有完全控制权。

在您购买了平台后，所属账户即为平台的管理员账户。可以使用“用户管理”功能，创建子用户，并给子用户分配管理员或操作员权限。不同用户权限如表 用户权限所示。

表 2-5 用户权限

用户角色	权限说明
管理员	拥有平台所有的权限，并进行用户管理，在平台添加子用户，以及对资源管理权限，包含有存储计算资源的购买和删除，自动扩缩容的策略配置权限。
操作员	拥有除用户管理、系统设置、设置商标、购买系统资源之外的所有权限。

### 说明

- 购买平台的账户为管理员账户，该账户不可被删除。
- 管理员可以创建子用户，并将子用户授权为管理员。子管理员同样具备“用户管理”功能。
- 对于系统级别的管理员用户（购买平台的账户），可以在平台右上角用户名中选择“用户管理”，在“操作”列中选择“安全设置”，重置其他用户的邮箱、手机和密码信息。
  - 单击“重置邮箱”后，可以选择清除邮箱或绑定新邮箱。
  - 单击“重置手机号”后，可以选择清除手机号或绑定新手机号。
  - 单击“重置密码”后，输入新密码和确认密码。用户密码重置成功后，首次登录成功需要再次修改密码。

## 2.4.2 创建子用户

1. 使用管理员账户登录医疗智能体平台。
2. 在右上角用户名中选择“用户管理”。

图 2-34 用户管理



3. 在用户管理页面，单击左上角“创建用户”，进入“创建用户”页面。
4. 填写用户信息，分配子用户权限。
  - a. 填写必填信息。必填信息包括“用户名”、“角色”、“密码”。“角色”分为管理员和操作员。权限区别请参见[表 用户权限](#)。

图 2-35 必填信息

\* 用户名:

\* 角色:

\* 密码:

- b. 选填信息。选填信息包括“邮箱”、“国际电话区号”、“手机号码”、“用户资源配额”。

图 2-36 选填信息

邮箱:

手机号:

用户资源配额:

5. 单击“确认”，完成子用户创建。

## 2.4.3 导入用户

EIHealth平台支持把华为云的IAM子用户导入至平台子用户。通过导入已有子用户，增加用户使用的便捷度，方便用户维护账号。

### 📖 说明

IAM用户导入只支持使用Domain用户导入。

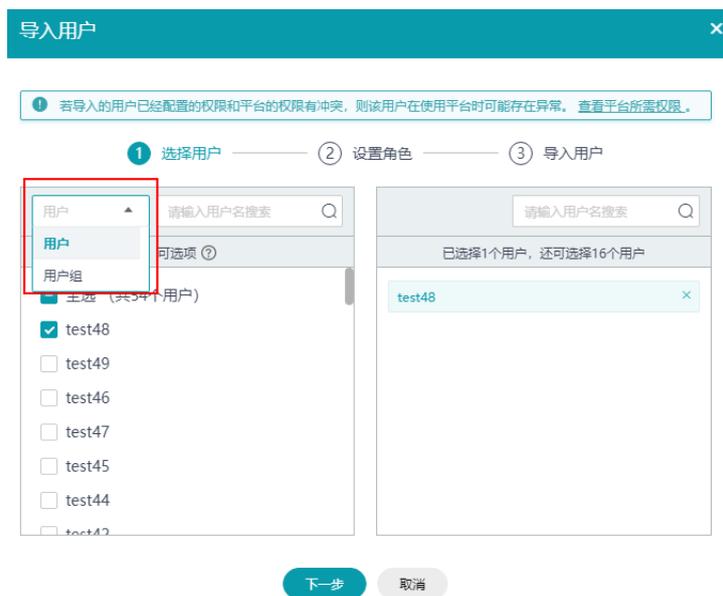
1. 使用管理员账户登录医疗智能体平台。
2. 在右上角用户名中选择“用户管理”。

图 2-37 用户管理



3. 在用户管理页面，单击“导入用户”，进入“导入用户”页面。
4. 在导入用户页面，可以选择“用户”或者“用户组”进行导入。

图 2-38 导入用户



### 说明

- 导入的IAM子用户需要具有管理控制台访问方式。
  - 导入用户时不能超出配额。如果超出配额，进行配额调整后，5分钟后生效。
  - 以用户组的方式导入时，若超出配额的部分会导入失败。
  - 以用户组的方式导入时，用户组里已经导入到平台的用户，不算统计个数。例如，用户组A里50个用户，10个已经导入平台，那么统计时，只会显示已选择40个用户。
  - IAM平台限制一个IAM用户不能加入超过10个用户组，而导入时还会再加入医疗用户组。如果IAM用户在导入平台之前就已经加入了10个用户组，则导入的时候会失败。
  - 用户数统计时会去重。例如，用户组A有50个用户，用户组B里有50个用，两者之间有10个用户重复，那么同时选择用户组A和B，统计时显示已选择90个用户。
  - 如果导入后用户状态显示异常，需联系技术支持处理。
5. 单击“下一步”，设置角色。
    - 角色：支持管理员和操作员两种角色，默认为操作员。权限描述可以参考[表 2-5](#)。
    - 用户资源配额：设置用户的个人资源配额。详细请参考[个人资源配额](#)。
  6. 配置完成后，单击“确定”。
  7. 等待导入成功后，单击“关闭”。可以在用户管理页面查看导入成功的用户信息。
- 导入的用户，不支持删除，只支持移除，移除后不影响该用户操作其他服务。

图 2-39 查看导入用户



## 📖 说明

医疗平台用户会在IAM中赋予以下细粒度权限，若该用户加入的其他IAM用户组有对应的deny权限，则会影响平台部分功能使用。

```
{
  "Version": "1.1",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "obs:bucket:GetBucketLocation",
        "obs:bucket:ListBucketMultipartUploads",
        "obs:object:GetObjectVersion",
        "obs:object:GetObject",
        "obs:object:GetObjectVersionAcl",
        "obs:object:DeleteObject",
        "obs:object:ListMultipartUploadParts",
        "obs:bucket:HeadBucket",
        "obs:object:AbortMultipartUpload",
        "obs:object:DeleteObjectVersion",
        "obs:object:GetObjectAcl",
        "obs:bucket:ListAllMyBuckets",
        "obs:bucket:ListBucket",
        "obs:object:PutObject"
      ]
    }
  ]
}
```

## 2.4.4 登录平台方式

- 管理员（购买平台的账户）
  - a. 登录医疗智能体管理控制台。
  - b. 选择账号登录，填写账号名和密码，登录平台。

图 2-40 登录界面

- 子用户（子管理员、操作员）
  - a. 登录医疗智能体管理控制台。

- b. 选择IAM用户登录，填写账号名、IAM用户名、密码，登录平台。
  - 账号名：与管理员（购买平台的账户）的账号名一致。
  - IAM用户名、密码：[创建子用户](#)时设置的用户名和密码。

图 2-41 IAM 用户登录界面



## 2.4.5 管理用户

### 删除子用户

管理员用户包含子管理员，均具备删除权限。但子管理员不可将管理员删除。删除用户只能删除来源为本平台的用户，如果来源为IAM，则只支持移除。

项目是存储数据、镜像、分析作业等的工作空间，执行删除用户操作时，只有在该用户名下没有项目时，才可以被删除。在用户的操作列，单击“删除”，删除对应的用户。

图 2-42 删除子用户



### 移除子用户

导入的用户，不支持删除，只支持移除。移除后不影响该用户操作其他服务。执行移除用户操作时，只有在该用户名下没有项目时，才可以被移除。在用户的操作列，单击“移除”，移除对应的用户。

图 2-43 移除子用户



## 安全设置

在用户管理页面，可以重置用户的邮箱、手机号和密码信息。在用户操作列，单击“安全设置”，重置用户的相关信息。

图 2-44 重置用户信息



## 配额设置

您可以对用户管理中的某个用户设置配额。在用户操作列，单击“配额设置”，设置用户资源配额。

图 2-45 用户配额设置



## 2.5 配额管理

为防止资源滥用，平台限制了各服务资源的配额，对用户的资源数量和容量做了限制。如果当前资源配额限制无法满足使用需要，您可以联系技术支持工程师申请扩大配额。

### 怎么查看我的配额

在平台右上角用户名中选择“配额管理”，查看配额。

图 2-46 配额管理

名称	资源名称	资源	已用配额	总配额
ElHealth	用户	用户总数	4	50
	项目	项目总数	6	100
		项目中的项目总数	--	20
		项目中的用户数	--	500
		项目中的项目数	--	500
	开发环境	项目中的开发环境数	--	10
	镜像	项目中的镜像数	--	500
	数据集	项目中的数据集实例数	--	500
		数据集数	--	500
		数据集存储数	--	1000000

## 2.6 系统设置

### 2.6.1 安全设置

当您需要对账号的安全信息进行设置时，可以在“安全设置”中进行操作。平台的任一用户均可对自己的邮箱、手机号码和密码进行修改。

在平台右上角用户名中选择“个人设置”，在“安全设置”中修改账号的安全信息。

图 2-47 安全设置

安全设置

邮箱   密码

手机号

#### 📖 说明

如果创建平台用户时，未填写邮箱或手机号，则在进行安全设置时，会提示设置邮箱或手机号。如果是最终租户，不支持修改密码。

### 2.6.2 商标设置

通过设置商标您可以自定义平台Logo和标语。设置商标功能只有管理员用户可进行操作。

在平台右上角用户名中选择“设置商标”，设置平台Logo和标语。首次修改时，需同时修改商标和标题。

图 2-48 设置商标

设置商标
✕

\* 商标 

\* 标题

上传图片要求

- 1.上传图片尺寸：高度32px，宽度<=60px。
- 2.图片类型为JPG/PNG
- 3.图片大小小于10KB

## 2.6.3 作业消息设置

### 设置作业保留条数

您可以在系统设置中设置作业最多保留条数。超过设置的数值后，系统将自动清除。

在平台右上角用户名中选择“系统设置”，设置作业的最多保留条数。已完成作业保留数量最少1万条，最多1000万条，默认设置为500万条。已有作业条数表示系统所有用户的作业总条数。

图 2-49 设置作业保留条数



### 设置消息保留条数

您可以在系统设置中设置消息最多保留条数。超过设置的数值后，系统将按照先后顺序自动清除。

在平台右上角用户名中选择“系统设置”，设置消息的保留条数。设置消息保留期功能只有管理员用户可进行操作。

消息通知最少1万条，最多1000万条，默认设置为500万条。已有消息条数表示系统所有用户的消息总条数。

图 2-50 设置消息保留条数



## 2.6.4 消息中心

消息中心用于所有用户的异步任务场景下的事件处理通知和系统通知，如项目删除、数据删除、数据导入、流程执行完成等，给予状态通知。

单击平台右上角图标，在弹出的“消息中心”中查看异步任务的操作记录和系统通知。

### 系统通知

在系统通知页签，可以查看系统通知的内容、时间。单击某个消息，可查看消息的详细内容。可以在该页面查看全部、未读和已读的消息。平台首页也会有未读消息提示。

图 2-51 平台首页查看消息通知



图 2-52 查看系统通知



- 支持标记为已读和批量标记为已读  
如果需批量标记已读，可以勾选多个未读的消息，单击“标记已读”。

图 2-53 批量标记已读



- 支持删除和批量删除  
如果删除单个消息，可以单击消息后面的  进行删除。

图 2-54 删除单个消息



如果批量删除消息，可以勾选多个消息后，单击“批量删除”。

图 2-55 批量删除消息



表 2-6 系统消息

消息类型	描述	查看权限
作业保留即将达到配额提醒	当前作业条数已经达到了配额的90%后，系统会消息提醒。	管理员
消息保留即将达到配额提醒	当前消息保留条数达到了配额的90%后，系统会消息提醒。	管理员
待删除项目提醒	待删除项目在删除前的3天，1天，系统会消息提醒。	项目内所有成员
平台即将升级通知	平台即将升级至X.X.X版本。	所有人
平台已升级通知	平台已升级至X.X.X版本。	所有人

## 操作记录

在操作记录页签，您可以按照“操作类型”、“资源类型”、“操作用户”、“项目名称”和“状态”对消息进行筛查。

例如，删除数据时，消息中心会显示数据所属的项目、资源的类型、删除操作的状态、操作人等信息。消息中心中呈现的内容请参见[表 消息类型](#)、[表 执行状态说明](#)。

图 2-56 操作记录

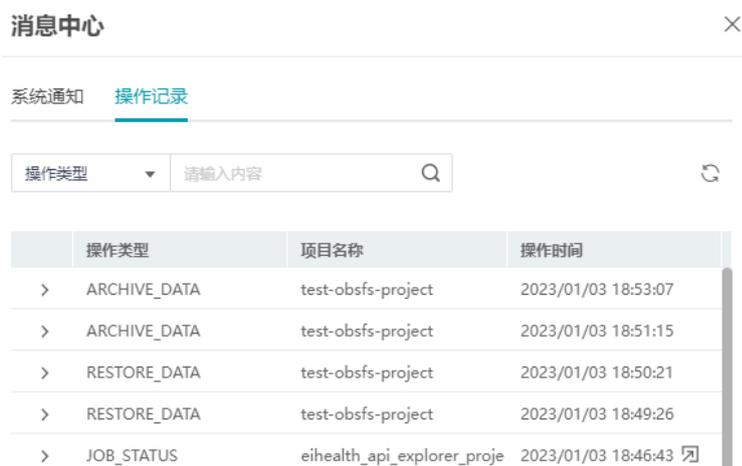


表 2-7 操作类型

操作类型	说明
PROJECT_DELETE	项目删除。某个项目开始删除、删除失败、删除成功时给予消息提示。
DELETE_DATA	数据删除。删除数据时给予消息提示，删除结果给予消息提示。
CLONE_DATA	数据拷贝。数据拷贝时给予消息提示，拷贝结果给予消息提示。
IMPORT_DATA	数据导入。数据导入时给予消息提示，导入结果给予消息提示。
ARCHIVE_DATA	归档数据。归档数据时给予消息提示，归档结果给予消息提示。
RESTORE_DATA	恢复数据。恢复数据时给予消息提示，恢复结果给予消息提示。
ARCHIVE_DATA_DELETE	删除归档数据。删除归档数据时给予消息提示，删除结果给予消息提示。
IMPORT_NETWORK_DATA	导入网络数据。导入网络数据时给予消息提示，导入结果给予消息提示。
SUBSCRIBE_DATA	订阅数据。订阅数据时给予消息提示，订阅结果给予消息提示。

操作类型	说明
DATABASE_IMPORT	导入数据库。导入数据时给予消息提示，导入结果给予消息提示。
JOB_STATUS	作业状态。作业的状态发生跳变时给予消息提示。
MESSAGE_CLEAN	消息清理。消息中心的消息总和超过设置值时，进行消息清理。

表 2-8 执行状态说明

执行状态	说明
SUCCEED	执行成功。
FAILED	执行失败。
PROCESSING	数据删除、导入等操作正在处理中。
START	项目删除操作开始执行。
RUNNING	分析作业运行中。
PENDING	分析作业等待处理中。
CANCELLED	取消作业。

### 📖 说明

- 对于“JOB\_STATUS”类型的消息，可单击图标，跳转至对应的分析作业详情页面。您可以在该页面查看详细的运行信息。
- 消息通知按照用户权限进行划分，只可以查看有权限访问的项目中的消息通知。

## 2.6.5 邮箱设置

### 设置消息发送邮箱

通过设置邮箱，发送平台的通知。设置邮箱功能只有管理员用户可进行操作。

1. 在平台右上角用户名中选择“系统设置”，设置邮件配置。
  - 服务器地址：邮箱开通SMTP功能时的服务器地址，不同邮箱开通SMTP方式不同，请使用搜索引擎查找邮箱开通SMTP方式。
  - 邮箱地址：填写邮箱地址，用于发送EIHealth平台的消息通知。
  - 用户名：邮箱的用户名，如果未修改过，默认为邮箱地址。
  - 密码：邮箱开通SMTP功能时生成的一段随机字符。
  - 消息头：发送邮件时，邮件标题。
  - 语言：选择中文或英文。

## 说明

删除邮箱后，将不再发送邮件通知。

图 2-57 邮件配置

系统发送邮件配置

服务器地址	smtp.163.com	邮箱地址	test@163.com
用户名	test@163.com	密码	
消息头	huawei	语言	中文 英文

测试 更新 删除

2. 邮箱配置完成后，单击“测试”，验证配置是否有误。配置成功后将在邮箱中收到测试通知。

图 2-58 邮箱测试通知



3. 收到邮箱测试通知后，单击“更新”，完成邮箱设置。

## 设置邮箱接收消息范围

在平台右上角用户名中选择“个人设置”，进入邮箱消息通知设置页面。设置邮箱消息的接收范围和接收类型。

- 接收范围
  - 仅自己接收操作：自己所执行的操作。如果选择此项，接收类型默认全选。
  - 全部接收：自己有权访问的项目中产生的操作，包含自己和其他项目成员所执行的操作。默认为全部接收。
  - 不接收：不接收操作通知。
- 接收类型
  - 数据操作：数据的复制、删除、导入等异步操作通知。
  - 作业进度：分析作业开始执行、执行成功通知。
  - 系统消息：项目删除、内存使用量告警、消息清理通知。

图 2-59 消息通知

邮件通知

接收范围

仅自己操作

全部接收

不接收

接收类型

数据操作

作业进度

系统消息

邮箱中接收到的消息通知如图2-60所示。

图 2-60 邮箱消息通知

Dear ei\_eihealth,

Project htt\_test has new messages you are following. Please confirm:

MessageType	ResourceType	ProjectName	ResourceName	Status	Operator	OperateTime
DATA_IMPORT	data	test	docking/rec/	SUCCEED	ei_eihealth	2021-03-02 19:42:25

Thank you for your support!

## 2.6.6 系统标签管理

在平台右上角用户名中选择“标签管理”。

系统标签用于设置项目、应用、流程、作业分类，系统管理员可以在系统标签管理中添加、删除、批量删除标签。

图 2-61 系统标签管理

名称	描述	操作
20221010		删除
预置数据集		删除
r46874959898565	预置数据集	删除
testbbb		删除
g-appy		删除

### 添加标签

1. 在“标签管理”页面，单击“添加”。
2. 在“添加标签”弹窗中，输入标签名称、标签描述。

标签名称长度为1~32，支持中文、字母、数字、空格、下划线和中划线，且不能以空格开头或者结尾。

图 2-62 添加标签



3. 单击“确定”，标签添加完成。系统最多可添加100个标签。

## 删除标签

标签删除后不可恢复，请谨慎操作。

1. 在“标签管理”页面，单击操作列的“删除”。
2. 在“删除标签”弹窗中，单击“确定”。

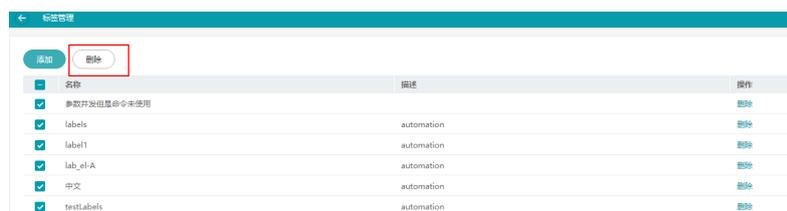
图 2-63 删除标签



## 批量删除标签

您可以一次选中多个标签，单击页面左上角的删除按钮即可。

图 2-64 批量删除标签



名称	描述	操作
参数开发组显命令未使用		删除
labels	automation	删除
label1	automation	删除
lab_e1-A	automation	删除
中文	automation	删除
testLabels	automation	删除

## 查找标签

在搜索框中输入标签名称关键字，单击  按钮进行查找。

图 2-65 查找标签



## 2.6.7 个人资源配额

在平台右上角用户名中选择“个人设置”，在页面最下方可以查看个人资源配额。

图 2-66 个人资源配额

### 个人资源配额

最大作业运行数	4
最大CPU占用核数	16
最大内存占用	64 GB
最大作业运行时间	1 天
最大创建项目数	2

## 个人资源配额设置方法

1. 登录系统管理员账号，在平台右上角用户名中选择“用户管理”，进入用户管理页面。在用户名最右侧的操作列，单击“配额设置”。

图 2-67 配额设置



2. 在“配额设置”页面，开启“用户资源配额”开关，填写“最大作业运行数”、“最大CPU占用核数”、“最大内存占用”、“最大作业运行时间”、“最大创建项目数”。

图 2-68 设置参数

配额设置: qh-test

用户资源配额

\* 最大作业运行数

\* 最大CPU占用核数

\* 最大内存占用  GB

\* 最大作业运行时间  天

\* 最大创建项目数

确定 取消

- 填写完成后，单击“确定”。

## 2.6.8 数据归档配置

设置平台数据归档的区域。

- 在账号下面选择“系统设置”。

图 2-69 选择系统设置



- 进入系统设置页面，设置数据归档配置区域。

图 2-70 配置数据归档区域



## 2.7 项目管理

### 2.7.1 项目简介

项目是EIHealth平台的一个工作空间，可以在项目中存储数据，上传镜像和创建分析作业。也可以将团队成员引入到项目中，并通过设置成员角色实现项目权限的划分。

项目管理是以项目为粒度对数据、分析作业、开发环境和镜像进行分组。以使用户通过项目进行资源的访问、共享和协作。您可以创建项目，并向其中上传数据、搭建流程、创建分析作业，对项目中的资产进行管理，同时不同项目间的资产可以通过“引用”进行分享和协作。

- [查看项目详情](#)
- [项目状态](#)

在“项目管理”页面，您可以查看项目的数量和总存储量。

图 2-71 项目管理



在“项目列表”中，展示了当前用户有权限访问的项目，以及项目大小、项目成员角色、所有者、状态、是否核心项目、创建时间、更新时间和可执行的操作。

图 2-72 项目列表

项目名称	大小	角色	所有者	状态	核心项目	创建时间	更新时间	操作
> zh-test	72.88GB	所有者	ei_ehealth...	可用	是	2022/05/07 09:20:45	2022/05/07 09:20:49	编辑   转移   分享   删除
> yjh-test-01	7.04GB	所有者	ei_ehealth...	可用	是	2022/05/16 11:26:11	2022/05/16 11:26:13	编辑   转移   分享   删除
> nyq-test	104.19GB	所有者	ei_ehealth...	可用	是	2022/04/22 15:58:41	2022/04/22 15:59:03	编辑   转移   分享   删除
> www-test-viewer	13.05GB	观察者	www-test-op	可用	是	2022/04/24 10:50:14	2022/04/24 10:52:45	编辑   转移   分享   删除
> www-test-uploader	7.63KB	上传者	www-test-op	可用	是	2022/04/24 10:49:56	2022/04/24 10:52:38	编辑   转移   分享   删除
> www-test-owner	143.16KB	管理员	www-test-a...	可用	是	2022/04/24 10:47:45	2022/04/24 10:55:57	编辑   转移   分享   删除
> www-test-developer	13.05GB	开发者	www-test-op	可用	是	2022/04/24 10:49:35	2022/04/24 10:52:31	编辑   转移   分享   删除
> www-test-admin	113.56GB	管理员	www-test-op	可用	是	2022/04/24 10:49:05	2022/04/24 10:52:23	编辑   转移   分享   删除
> user-demo	12.96GB	所有者	ei_ehealth...	可用	是	2022/04/24 10:50:18	2022/05/07 16:06:13	编辑   转移   分享   删除
> test123	12.77GB	所有者	ei_ehealth...	可用	是	2022/04/22 10:20:05	2022/05/07 16:38:38	编辑   转移   分享   删除

### 查看项目详情

项目详情提供了项目中数据大小、成员、创建时间等信息，您可以通过以下两种方式查看项目信息。

● 查看项目概览

单击项目名称左侧 > 图标，或双击项目行中的空白处，展开项目概览页面。您可以在该页面查看项目的描述、标签、创建时间、更新时间、是否核心项目、数据大小和成员信息。

图 2-73 项目概览



● 查看项目详细信息

单击项目名称，进入项目“设置”页面。您可以在该页面，查看项目的数据存储量、作业总量、流程数量和应用数量，修改项目是否核心项目。同时，可以查看项目的基本信息、数据控制信息和成员信息。

图 2-74 项目详细信息



## 项目状态

- 可用：项目当前状态正常。项目创建完成后进入“可用”状态，可用状态的项目允许被冻结、删除、转移和分享，执行此类操作需要该项目的成员拥有相关权限。
- 冻结：项目当前不可用。处于“冻结”状态的项目，用户无法进入该项目查看项目的开发环境、流程等历史运行情况。冻结的项目可以通过解冻操作重新激活。
- 删除中：项目删除中。需先冻结项目，才能执行删除操作。删除的核心项目，将进入“待删除项目”列表中。待删除项目会保留7天，6天内您可以将项目恢复成可用状态，最后一天不支持恢复。7天后，项目将自动删除，删除后不可恢复。非核心项目支持立即删除。

图 2-75 项目状态



## 2.7.2 创建项目

您可以在“项目管理”页面创建一个新的项目。

1. 在“项目管理”页面单击“创建项目”。
2. 配置项目信息。

表 2-9 参数说明

参数	说明
项目名称	项目名称长度限制3-45，以小写字母数字开头结尾，全文包含数字、小写字母、下划线、中划线。
核心项目	如果设置该项目为核心项目，不支持立即删除，会进入待删除项目列表。删除项目需要等待7天保留期，到期后系统自动删除。 <b>说明</b> 一旦设置为核心项目，不可变为非核心项目，非核心项目支持变为核心项目。
标签	设置项目标签。
描述	设置项目描述。
数据保护策略	数据保护策略介绍请参见 <a href="#">数据控制与数据审计</a> 。

图 2-76 创建项目

**新建项目** ✕

创建项目来管理数据、工具、数据库、作业。

\* 项目名称

\* 核心项目  ?

标签

最近常用标签：  
客户1 smallMolecularC...  
peptide-docking

描述

\* 数据保护策略

\* 分享  开启  关闭 关闭分享后，项目数据不允许分享给其他项目，包括...

\* 下载  开启  关闭 关闭下载后，项目数据不允许下载到本地。

\* 删除  开启  关闭 关闭删除后，项目数据不允许手动删除。

0/1024

3. 单击“确认”，创建一个新的项目。

### 📖 说明

- 项目的创建者默认拥有项目的完整权限，同时项目可以分享给其他用户，并限定其他用户的访问权限。项目角色为项目粒度权限控制，同一用户在不同的项目上可能拥有不同的角色。
- 创建的项目配额请参见[配额管理](#)进行查询。详细添加项目成员并分配角色的方法请参见[添加项目成员](#)。

## 2.7.3 项目成员和权限

医疗智能体以项目为粒度对数据、分析作业、开发环境和镜像进行分组。以使用户通过项目进行资源的访问、共享和协作。

- [添加项目成员](#)
- [移除、修改项目成员](#)
- [成员角色和权限](#)

### 添加项目成员

#### 前提条件

- 平台管理员首先通过“用户管理”功能添加平台用户，才能将该用户添加至项目中。  
创建用户的详细方法请参见[创建平台用户](#)。
- 存在一个创建好的项目。

#### 操作步骤

添加项目成员有两种不同的方法，请任选其中一种方法操作。

- 方法一
  - a. 单击项目名称，进入项目“设置”页。
  - b. 单击“添加”，添加成员。

图 2-77 添加成员



- c. 输入已添加至平台的用户的全称。

图 2-78 输入用户名全称



- d. 单击“添加”，设置用户角色。详细成员角色介绍请参见表2-10。

图 2-79 设置成员角色



- e. 单击“确认”，将用户添加至项目中。
- 方法二
    - 在项目列表中，单击“操作”列“分享”。

图 2-80 分享项目



- b. 输入已添加至平台的用户的全称。

图 2-81 输入用户名全称



- c. 单击“添加”，设置用户角色。详细成员角色介绍请参见表2-10。

图 2-82 设置成员角色



- d. 单击“确认”，将用户添加至项目中。

## 移除、修改项目成员

### 操作步骤

1. 单击项目名称，进入项目“设置”页。
2. 单击“操作”列“修改”，重新设置项目成员的角色。  
详细项目成员角色和权限介绍请参见[成员角色和权限](#)。

图 2-83 移除、修改项目成员



3. 选中成员，单击“移除”，将成员从项目中移除。选中多个成员，可进行批量移除，一次最多选中五个成员。

## 成员角色和权限

在项目中，成员可以划分为5种不同的角色。详细角色介绍请参考[表 项目成员角色](#)，详细角色权限请参考[表 角色权限](#)。

表 2-10 项目成员角色

角色	说明
所有者	项目所有者具备该项目完整的权限，所有者允许删除项目。同时，可以将项目转移给其他用户。
管理员	项目管理员具备除了删除、冻结、转移、解冻、恢复外其他的完整项目权限。
开发者	项目开发者可在项目内执行完整的操作，如Notebook、流程管理等，但无项目分享等权限。
上传者	项目数据上传者，可通过EIHealth平台或命令行工具上传数据。无法执行流程和Notebook等操作。
观看者	项目查看人员仅具备项目查看权限，无项目操作权限。

表 2-11 角色权限

成员角色		所有者	管理员	开发者	上传者	观看者
项目管理	状态管理	√	-	-	-	-
	创建	√	√	√	√	√
	修改	√	√	-	-	-
	查询	√	√	√	√	√
	分享	√	√	-	-	-

镜像管理	上传、下载、删除	√	√	√	-	-
	发布	√	√	-	-	-
	查看	√	√	√	√	√
数据管理	上传、导入	√	√	√	√	-
	发布	√	√	-	-	-
	下载	√	√	√	-	-
	删除	√	√	√	-	-
	查看	√	√	√	√	√
流程管理	创建、更新、删除、导入、订阅	√	√	√	-	-
	发布	√	√	-	-	-
	查看、导出、下载	√	√	√	√	√
应用管理	创建、更新、删除、导入、订阅	√	√	√	-	-
	发布	√	√	-	-	-
	查看、导出、下载	√	√	√	√	√
作业管理	启动、停止、取消、重试、更新、删除	√	√	√	-	-
	查看、导出、下载	√	√	√	√	√
自动作业管理	启动、停止、创建、更新、删除	√	√	√	-	-
	查看	√	√	√	√	√

Nextflow	启动、停止、重试、创建、更新、删除	√	√	√	-	-
	查看	√	√	√	√	√
开发环境	创建、导入	√	√	√	-	-
	查看	√	√	√	√	√
	启、停	√	√	√	-	-
数据库	创建、导入、删除	√	√	√	-	-
	查看	√	√	√	√	√
药物虚拟筛选	创建、修改、删除	√	√	√	-	-
	查看	√	√	√	√	√

### 说明

- 项目列表中只显示您所在的项目。
- 一个用户可以是多个项目的成员。

## 2.7.4 项目管理常用操作

对项目执行冻结、解冻、转移、分享和删除操作时，需要项目成员具备相应的权限，不同类型成员所拥有的权限请参见[成员角色和权限](#)。

- 冻结：冻结项目。一个项目可能由于某些原因暂时中止，此时可以将项目暂时冻结，待下次重新启动时再激活。处于“冻结”状态的项目，用户无法进入该项目查看项目的开发环境、流程等历史运行情况，当前正在运行的分析作业会继续执行直到完成。冻结的项目可以通过“解冻”操作重新激活。
- 解冻：解冻项目。
- 转移：将项目转移给其他用户。“转移”只能由项目所有者发起，并只能转移给管理员，项目转移后，原用户角色由所有者转变为管理员。
- 分享：将项目分享给其他用户，并设置项目角色。项目分享后，项目将出现在被分享者的项目列表中。
- 删除：删除项目。需先冻结项目，才能执行删除操作。
- 恢复：删除的项目在删除后6天内都可以恢复成“可用”或者“冻结”状态。

图 2-84 项目管理



## 分享项目

### 前提条件

- 平台管理员首先通过“用户管理”功能添加平台用户，才能将该用户添加至项目中。  
创建用户的详细方法请参见[创建平台用户](#)。
- 已创建一个项目。

### 操作步骤

- 单击“操作”列“分享”。

图 2-85 分享项目



- 输入已添加至平台的用户名称。

图 2-86 输入用户名全称



- 单击“添加”，设置用户角色。详细成员角色介绍请参见[表2-10](#)。

图 2-87 设置成员角色



- 单击“确认”，分享项目。

## 转移项目

“转移”只能由项目所有者发起，转移给项目管理员。项目转移后，原用户角色由所有者转变为管理员。

### 操作步骤

- 单击“操作”列“转移”。

图 2-88 转移项目



2. 选择项目交接的用户。
3. 单击“确认”，转移项目。

## 设置项目存储容量

平台支持项目管理员以上角色，配置项目的最大存储量。

1. 单击项目名称，在项目页面选择“设置”。
2. 在项目详情页面，单击“已用存储量”后面的 。

图 2-89 设置存储容量



3. 开启容量限制开关后，设置最大存储量。

图 2-90 容量限制



4. 设置成功后，单击“确定”。

### 说明

项目存储量15分钟刷新一次，如果设置了项目最大存储量，项目数据达到最大存储量后，数据上传、复制、导入、执行的作业、notebook的使用会失败。

## 2.8 数据管理

### 2.8.1 数据管理简介

EIHealth平台使用对象存储服务（OBS）存储原始数据、流程执行中间数据和执行结果数据。数据按**项目**维度进行隔离和划分，从项目角度进行数据的管理，不同项目的数据可以通过“导入数据”实现跨项目使用。您可以在“项目管理”页面，“数据”页签中，完成数据的添加、导入、归档、复制、删除等操作。

#### 📖 说明

- 数据总条数≤10000，页面上可以指定跳转具体页码。
- 数据总条数>10000，通过“下一页”和“上一页”进行页面切换，并且按时间排序只能对当前页面进行排序，页面搜索为按前缀搜索。

### 使用合规数据

您添加的数据属于您的内容，您应对您的内容的合法合规性负责，建议您根据对您适用的法律法规，进行相应的脱敏、匿名、加密或其他符合要求的处理或采取相应保护措施，当您在复制或者引用数据的时候，需审视是否存在隐私数据，请谨慎操作。

### 上传数据方式

在开始执行分析作业前，请先上传待分析的原始数据。不同的上传方法对数据大小要求不同，您可以参考[表 上传数据方式](#)选择相应的数据上传方式。

表 2-12 上传数据方式

上传数据方式	说明
“数据”页面上传	通过“数据”页面上传数据，支持上传最大为1GB的单个文件。 数据上传方法请参见 <a href="#">“数据”页面上传</a> 。
使用命令行工具上传	命令行工具（eihealth-toolkit）配套EIHealth平台，提供数据、应用、流程和作业资源的管理和使用，支持上传最大为48.8TB的单个文件。 数据上传方法请参见 <a href="#">命令行工具概述</a> 。

### 2.8.2 添加数据

在添加数据前，您可以创建存储数据的文件夹，将数据存储在对应的文件夹中。如果已有存储数据的文件夹，可以直接添加数据。

#### （可选）新建文件夹

1. 单击项目名称，并选择“数据”。

图 2-91 数据管理



- 单击“新建文件夹”，创建文件夹。

**说明**

文件夹名称命名规则：

- 支持创建单个文件夹和多层级的文件夹。
- 单个斜杠 (/) 表示分隔并创建多层级的文件夹。
- 单个文件夹名称不能包含以下字符：\ / : ; \* ? < " > |。
- 文件夹的绝对路径总长度不能超过1023字符。
- 文件夹名称不能以英文句号 (.) 或斜杠 (/) 开头或结尾。
- 不能包含两个以上相邻的斜杠 (/)。
- 文件夹名称不能以中划线 (-) 开头。
- 多层文件夹不能超过45层。
- 数据列表按照先文件夹后文件的顺序排列。文件夹默认为首字母排序，文件可按照时间顺序排序，但始终在文件夹后面显示。

## 上传数据

- 选择数据需要上传的文件夹，并单击“添加数据”或者直接单击“添加数据”。
- 在添加数据页面，选择“上传”，上传待分析数据。
- 数据上传成功后，单击“确定”。

**说明**

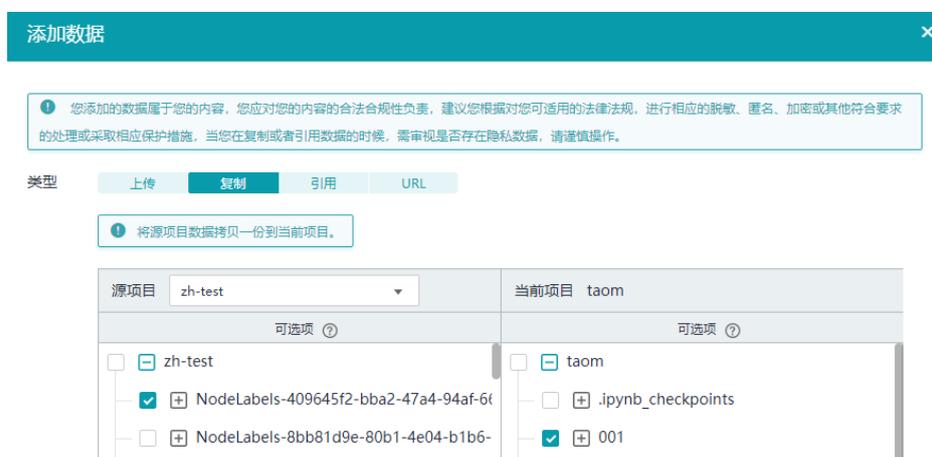
上传文件大小不能超过1GB。

## 复制数据

将其他项目的数据，复制到本项目中，并可在本项目中操作该数据。

- 单击“添加数据”，类型选择“复制”。
- 选择需要复制的项目数据，选择导入数据的路径。
  - 您可以选中多个文件夹进行批量复制。
  - 复制数据时可单击 图标创建文件夹。

图 2-92 复制数据



3. 单击“确认”，复制其他项目中的数据至本项目。

## 引用数据

将其他项目或OBS桶中的数据，引用到本项目，不可在本项目中操作该数据。

1. 单击“添加数据”，类型选择“引用”。

图 2-93 引用数据



2. 选择需要引用的项目以及项目中的数据，或者选择待引用的OBS桶路径，先选择OBS桶所在区域，再选择OBS桶名称，支持选择不在同一区域的OBS桶。
3. 单击“确定”，引用其他项目中的数据至本项目。

引用的数据和项目将显示在左侧的数据列表中。

图 2-94 引用的数据



## 📖 说明

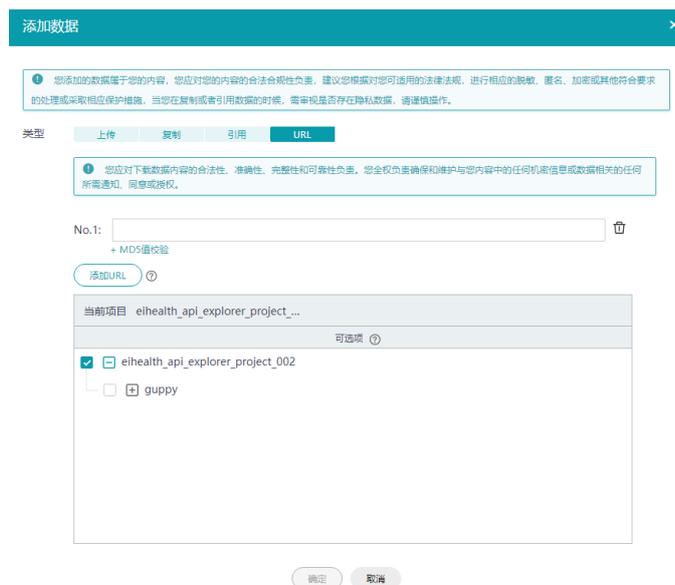
- 引用OBS类型数据时，如果数据在OBS中的存储类型为“归档存储”，则将该数据引用过来后，该数据不能用于创建分析作业、数据库导入数据，并不可下载、归档。
- 引用外部OBS桶的数据时，在页面左侧数据栏中，将显示出完整的外部桶目录结构，对于非引用路径的数据，不支持数据归档操作。例如，桶中有文件，路径为a/b/c，引用c文件，c文件的父目录也会显示在左侧数据栏中，但目录a，目录b 不支持数据归档。
- 平台系统管理员在自己的所有者、管理员、开发者、上传者项目可以引用OBS类型数据。平台系统管理员在自己的所有者、管理员、开发者、项目可以解除引用OBS类型数据。其他角色的用户仅能使用引用进来的OBS类型数据。

## 使用 URL 导入数据

使用URL方式将公网http/https/ftp链接的数据或者OBS中的数据导入。

1. 单击“添加数据”，类型选择“URL”。
2. 单击“添加URL”，填写URL地址。
3. 选择需要导入数据的文件夹。
4. 单击“确定”，导入数据。

图 2-95 使用 url 导入数据



## 2.8.3 发布数据

您可以将数据发布到资产市场，供其他项目订阅和使用。

发布资产前，需要先设置商标。详细可参考[商标设置](#)。

1. 单击项目名称，进入项目管理页面，选择“数据”页签。
2. 选择需要发布的数据或者文件夹，单击“操作”列的“更多>发布”。

图 2-96 发布数据



3. 在发布页面，填写发布数据的相关信息。

表 2-13 发布数据参数说明

参数	说明
名称	发布数据的名称。名称长度必须为1-128个字符，只支持字母，数字，中划线和下划线。
展示名称	发布成功后在组织共享中展示的名称。若不填写，展示的名称跟发布数据的名称保持一致。发布成功的数据不支持修改展示名称。
版本	发布数据的版本。 版本长度必须为1-64个字符，只支持字母，数字，中划线，下划线和点号。
类型	数据。
标签	设置发布数据的标签。 最多支持设置5个标签。
短描述	为发布的数据设置的简要描述。
图片	展示在资产市场的图片。要求如下： <ul style="list-style-type: none"> <li>● 格式：svg,png,jpg,jpeg</li> <li>● 尺寸：16 * 9</li> <li>● 大小：不超过20KB</li> </ul>
描述	针对此数据的相关描述。

4. 单击“立即发布”。

系统跳转到“我的发布”页面，可以查看发布的数据状态。当数据发布成功后，可以在“资产市场>组织共享”中查看发布的数据。

#### 📖 说明

- 已发布的数据不能直接编辑和删除，需下线后才能编辑和删除。
- 发布中的数据可以取消，取消后状态为发布失败。
- 已发布的数据可以订阅和下线。
- 同一个发布者可以发布同名称，同类型的数据，若版本不同，则合并。用户在使用时，可以选择对应的版本号进行订阅或者下线。若版本相同，则更改版本后发布。
- 已发布的数据，可以在“资产市场>我的发布”中进行下线。下线后已经订阅的数据不可以使用。

## 2.8.4 数据管理常用操作

- [复制数据](#)
- [解除引用](#)
- [下载数据](#)
- [禁止/允许删除数据](#)
- [删除数据](#)
- [恢复数据](#)

执行数据管理类操作需要项目成员具备相应的权限，详细的权限介绍请参见[项目成员和权限](#)。

数据文件的名称，不可以含有特殊字符。如果文件名包含特殊字符，将不支持下载，可通过去除文件名中的特殊字符方式解决。

### 查看数据作业

数据拷贝、数据删除、数据下载、数据归档、恢复数据等操作均为系统的异步任务，您可以在“消息中心”中查看任务的执行状态。同时，平台针对该类任务支持您在“作业 > 数据作业”中查看任务的执行信息。在数据作业页面，支持通过作业类型、创建者、状态、创建时间、完成时间来进行搜索。

数据作业支持取消、删除、重试操作，三种操作均支持批量操作。

**取消：**一个数据作业中，单个数据操作执行过程为原子性操作，即要么不执行，要么需要执行过程不能中断。在执行完成后，“取消”操作才能生效。

**删除：**运行中的任务不支持删除，只可删除运行结束的数据。

**重试：**可对执行失败的任务从失败位置进行重试。

图 2-97 数据作业

作业名称	类型	状态	创建者	创建时间	完成时间	运行时间	操作
NetworkDataImport-20220620083108	IMPORT_NETWORK_DATA	已取消	ei_ehealth_ll	2022/06/20 16:31:08	2022/06/20 16:31:48	00:00:40	删除   重试
DataDelete-20220620082633	DELETE_DATA	成功	ei_ehealth_ll	2022/06/20 16:26:33	2022/06/20 16:26:34	00:00:01	删除
NetworkDataImport-20220620023140	IMPORT_NETWORK_DATA	成功	ei_ehealth_ll	2022/06/20 10:31:40	2022/06/20 10:35:04	00:03:24	删除
DataDelete-20220620022745	DELETE_DATA	成功	ei_ehealth_ll	2022/06/20 10:27:45	2022/06/20 10:27:49	00:00:04	删除
DataDelete-20220620022736	DELETE_DATA	成功	ei_ehealth_ll	2022/06/20 10:27:36	2022/06/20 10:27:38	00:00:02	删除
NetworkDataImport-20220614123621	IMPORT_NETWORK_DATA	成功	ei_ehealth_ll	2022/06/14 20:36:21	2022/06/14 20:36:24	00:00:03	删除
NetworkDataImport-20220614123220	IMPORT_NETWORK_DATA	成功	ei_ehealth_ll	2022/06/14 20:32:20	2022/06/14 20:36:09	00:03:49	删除
NetworkDataImport-20220614122825	IMPORT_NETWORK_DATA	已取消	ei_ehealth_ll	2022/06/14 20:28:25	2022/06/14 20:28:49	00:00:24	删除   重试
DatabaseImport-93713642-a35f-48a1-80a5-f69b5	DATABASE_IMPORT	已取消	ei_ehealth_ll	2022/06/13 15:54:44	2022/06/13 15:55:10	00:00:26	删除   重试
DatabaseImport-93713642-a35f-48a1-80a5-f69b5	DATABASE_IMPORT	成功	ei_ehealth_ll	2022/06/13 15:48:04	2022/06/13 15:48:12	00:00:08	删除

### 复制数据

在“数据”页面，使用“复制”，将本项目的某一个文件、文件夹复制到其他文件夹中。

1. 单击“复制”。

图 2-98 复制数据



2. 选择需要复制的文件或文件夹，以及数据存储路径，单击“确认”，复制数据。

例如，将“test.zip”文件复制到“abc”文件夹中。复制数据时可单击  图标创建文件夹。

图 2-99 复制数据



## 解除引用

将已经引用的数据解除引用关系。有以下两种方式，可以实现解除数据的引用。

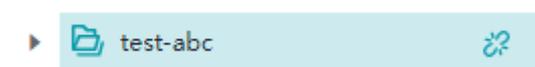
- 单击“操作”列的“解除引用”，解除数据引用关系。

图 2-100 解除引用



- 鼠标指向左侧数据列表，选择需要解除引用的数据。单击右侧  图标，解除引用关系。

图 2-101 解除引用



## 下载数据

### 须知

下载数据操作会产生流量费用，计费方式为按需计费，计费详情请参考[OBS数据下载费用](#)。

以下操作步骤是在“数据”页面下载数据至本地。您也可以使用[命令行工具](#)实现数据的下载。

1. 展开数据文件夹，选择待下载的数据。
2. 单击“操作”列“下载”。
3. 单击鼠标右键，选择“链接另存为”，下载数据。

图 2-102 下载数据



## 禁止/允许删除数据

您可以对某个数据设置禁止删除。设置禁止删除后，该目录只能添加数据，不支持删除数据。也可以通过“允许删除”取消禁止删除设置。

图 2-103 开启禁止删除数据



### 说明

支持设置最多15个数据的禁止删除状态。

如果平台或者父目录设置禁止删除状态，则对应的数据根路径或者子路径均为禁止删除状态，子路径不支持同时设置允许删除状态；如果平台或者父目录设置允许删除，则之前子路径设置的禁止删除权限失效，全部子路径允许删除。

## 删除数据

以下操作步骤是在“数据”页面删除数据。您也可以使用[命令行工具](#)实现数据的删除。

- 删除数据  
单击“操作”列的“更多>删除”，删除数据。

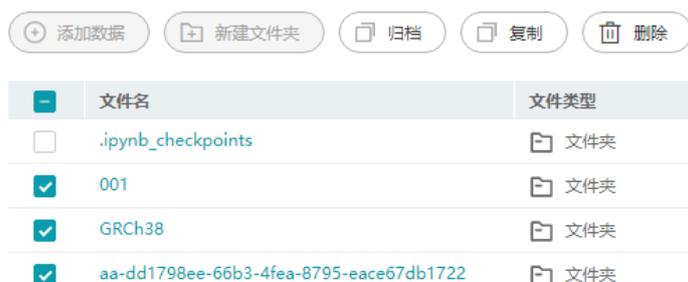
图 2-104 删除数据



● 批量删除

勾选文件名左侧  图标，单击“删除”，删除数据。

图 2-105 批量删除



📖 说明

批量删除时，若选择的文件包含禁止删除的文件或文件夹，则这些文件或文件夹会删除失败，其他文件/文件夹删除正常。

## 恢复数据

平台支持数据的恢复机制，您可以将归档的核心数据进行恢复，避免造成损失。

数据归档成功后，可以在“归档”页面查看归档记录，并进行恢复、删除操作。归档恢复时，您可以将数据恢复至本项目或有权限的其他项目。

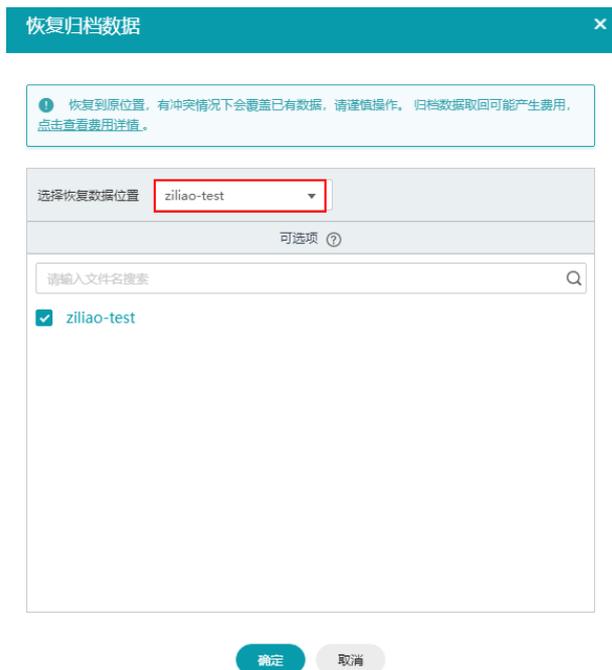
1. 数据归档成功后，选择“归档”页签。

图 2-106 选择归档页签



2. 选择需要恢复的数据，单击操作列“恢复”。在恢复归档数据页面，选择需要恢复数据的位置和数据。

图 2-107 选择恢复数据的位置



3. 单击“确定”。

## 2.8.5 归档数据

平台支持数据的归档机制，您可以将核心数据进行归档，避免误删造成损失。

1. 在“数据”页面单击“归档”，进入归档数据页面。

图 2-108 选择归档



2. 填写归档的参数信息。

表 2-14 参数说明

参数	说明
名称	设置归档数据的名称。归档名称长度为1-100。
描述	归档数据描述信息。
归档类型	选择归档数据存储类型。默认为标准存储。 <ul style="list-style-type: none"> <li>标准存储：适合高性能，高可靠，高可用，频繁访问场景。</li> <li>归档存储：适合长期存储，平均一年访问一次。</li> </ul> <b>说明</b> 归档存储的费用为单独计费。

参数	说明
删除原数据	是否删除原数据。如果选择删除原数据，系统会在归档完成后统一删除已归档数据。归档后的数据可在归档中心查看或恢复。 在归档过程中请不要对归档的文件或者文件夹进行添加、更改操作，以免造成您的数据损失。
选择归档数据	选择该项目下需要归档的数据。

3. 单击“确认”，进行归档。

归档操作将记录在“归档”页面，您可以在该页面查看归档记录，并进行归档恢复、删除操作。归档恢复时，您可以将数据恢复至本项目或有权限的其他项目。

图 2-109 归档数据

归档ID	归档路径	归档类型	归档次数	归档路径	描述	开始时间	结束时间	大小	操作人员	操作
456	7f8b259e-7a89-4764-ac31-72ca389023ac	备份	0	guppy/PA22933...	备份	2023/01/05 09:2...	2023/01/05 09:2...	89.24MB	ej_ehealth_jw@...	恢复 删除
123	af6e4540-420b-4632-b659-9032a3d238c	备份	0	ligands.txt	--	2023/01/04 19:1...	2023/01/04 19:1...	4.91KB	ej_ehealth_jw@...	恢复 删除
test123	30a50a3e-e031-4884-8887-191ea3682185	备份	10	guppy/guppy@...	--	2022/12/28 10:1...	2022/12/28 10:1...	89.25MB	ej_ehealth_jw@...	恢复 删除

说明

- 执行归档操作时，归档的对象，其最深目录下的对象路径长度不能超过987，否则会归档失败。例如，选择项目中的 a/目录进行归档，a/目录下最深一级的文件 a/xxxx/xxx./.../obj的总长度不能超过987。
- 执行归档操作时，若包含禁止删除的文件或文件夹，并打开了删除原数据开关，则被设置为禁止删除的文件或文件夹会删除失败，其他文件或文件夹删除正常。
- 使用obsutil上传数据，必须加full controle，否则会导致归档失败。
- 归档作业取消后，不支持重试。
- 选择归档类型为归档存储后，对应的归档数据至少要保存90天，如果在90天内删除归档数据，需要支付不足90天的惩罚费用。详细可参考[不同类型的归档，费用分别是多少](#)。

## 2.8.6 数据控制与数据审计

### 数据保护策略

项目内的数据支持精细化的权限控制，可对数据分享、下载、删除进行设置。您可以在项目的“设置”页面设置数据权限。数据权限仅可以有项目所有者设置。

- 分享：关闭分享后，项目内数据不允许分享给其他项目，包括拷贝、引用两种方式。
- 下载：关闭下载后，项目内数据不允许下载至本地。
- 删除：关闭删除后，项目内数据不允许通过平台、命令行工具删除。

图 2-110 数据保护策略

数据控制



## 数据审计

平台通过云审计服务（CTS）提供操作记录的收集、存储和查询，审计操作可以设置导出用户的写操作（默认）和读操作（可选），保存周期可以设置需要保存的审计日志的保存周期，下载按钮可以下载最近7天内最新的1万条数据审计日志，查看按钮可以查看及下载保存的审计日志。

通过审计日志可用于支撑安全分析、合规审计、资源跟踪和问题定位等常见应用场景。您可以在项目的“设置”页面下载最近7天的数据审计日志，其他操作的审计日志请登录云审计服务控制台查看。

图 2-111 数据审计



### 2.8.7 命令行工具概述

医疗智能体命令行工具（eihealth-toolkit）是配套EIHealth平台，为EIHealth平台各功能组件提供命令行管理工具。借助此工具，可以辅助您对EIHealth平台项目中数据、应用、流程和作业资源进行管理和使用。

eihealth-toolkit具备灵活性高、扩展性强：

- 单一可执行文件，方便拷贝与安装。
- 支持多操作系统，包括Windows、Linux和MacOS，满足不同开发者需求。

使用命令行工具执行数据操作的详细步骤如下所示，详细的命令行介绍请参见[医疗智能体命令行工具](#)。

- [步骤1：下载eihealth-toolkit](#)
- [步骤2：安装eihealth-toolkit](#)
- [步骤3：初始化配置](#)
- [步骤4：上传数据](#)
- [步骤5：下载数据](#)
- [清理历史命令](#)

### 前提条件

[获取认证信息](#)（AK/SK，区域名称，平台ID）。

### 步骤 1：下载 eihealth-toolkit

针对不同操作系统，eihealth-toolkit下载地址如下所示。

表 2-15 下载列表

支持平台	下载地址
Windows 64位	<a href="#">health-windows-x86_64.zip</a> 、 <a href="#">health-windows-x86_64.zip.sha256</a>
Linux ARM 64位	<a href="#">health-linux-aarch64.tar</a> 、 <a href="#">health-linux-aarch64.tar.sha256</a>
Linux AMD 64位	<a href="#">health-linux-x86_64.tar</a> 、 <a href="#">health-linux-x86_64.tar.sha256</a>
macOS	<a href="#">health-macOS-x86_64.tar</a> 、 <a href="#">health-macOS-x86_64.tar.sha256</a>

### 📖 说明

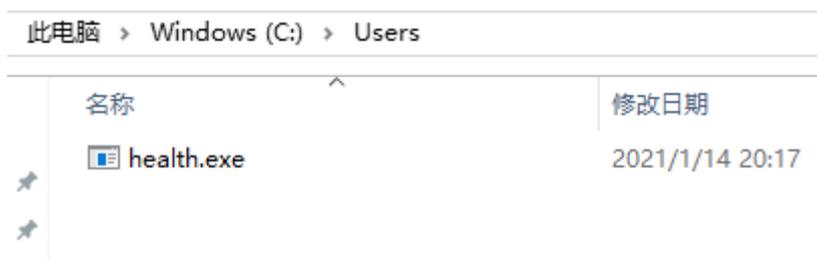
本页面命令行工具下载后，在使用时，需用到您注册华为账号并开通华为云时提供的用户名等信息，用于登录并操作EIHealth平台的项目、数据等资产。这些信息的处理将遵循您已接收的《[华为云用户协议](#)》及《[隐私政策声明](#)》约束。

## 步骤 2：安装 eihealth-toolkit

本示例中以Windows系统为例，介绍安装命令行工具的方法。

1. 下载Windows版本的客户端，得到health.exe文件，health文件无需安装，放置在任一文件夹中即可。

图 2-112 下载命令行工具



2. 使用win键+R，输入cmd打开windows的cmd窗口。进入工具所在的目录，输入health命令，即可使用。

如果cmd窗口显示目录不是health文件所在目录，请使用cd命令切换路径。例如，切换至D盘：

```
cd /d d:
```

图 2-113 客户端

```
C:\Users>health
A simple command client for ei health.

Usage:
health [flags]
health [command]

Available Commands:
cd          change project directory
config     add configuration information
cp         copy project objects
create     create a app,workflow,job...
delete     delete a app,workflow,job...
download   download objects from project directory to local directory
edit       edit a app,workflow...
get        get project, app, workflow or tool
help      Help about any command
ls         list project objects
mkdir     create new project directory
pwd       print current work directory
rm        delete project objects
switch    switch to a project
upload    upload objects from local directory to project directory

Flags:
-h, --help  help for health

Use "health [command] --help" for more information about a command.
```

 说明

- 使用Linux版本命令行工具时，您需要在本地搭建Linux环境，并将下载的health文件放置所需的目录下，例如：
  - 如果当前目录为health所在目录，可以使用./health命令执行命令行工具。
  - 如果当前目录不是health所在目录，需要使用绝对路径。如当前目录为/opt，假设health存放在/root/health-toolkit/下，需要指定/root/health-toolkit/health路径进行使用。
- 如果无法运行，使用chmod 755 health命令设置执行权限。

### 步骤 3：初始化配置

在使用命令行工具前，需要初始化配置信息，通过config命令对eihealth-toolkit进行初始化配置。本节以Windows为例介绍配置过程，Linux和macOS环境配置过程相同。

- 命令结构  
执行health config add命令配置AK、SK、region、platform-id信息，获取方法请参见[获取认证信息](#)。

```
health config add [flags]
```

表 2-16 参数说明

参数	简写	是否必选	说明
--domain-name	-d	是	与管理员（购买平台的账户）的账号名一致。
--user-name	-u	是	子用户的用户名。 管理员（购买平台的账户）登录时，user-name和domain-name一致。
--password	-w	是	密码。
--ak	-a	是	AK(Access Key ID)：访问密钥ID。

参数	简写	是否必选	说明
--sk	-s	是	SK(Secret Access Key): 与访问密钥ID结合使用的密钥。
--region	-r	是	服务区域名称。
--platform-id	-i	是	平台ID, 获取方法请参见 <a href="#">获取认证信息</a> 。
--iam-endpoint	-m	否	IAM终端节点名称, 请在 <a href="#">地区与终端节点</a> 中获取。
--health-endpoint	-e	否	EIHealth终端节点名称, 请在 <a href="#">地区与终端节点</a> 中获取。
--swr-endpoint	-t	否	SWR镜像仓库地址。 获取方式: 1. 登录容器镜像服务管理控制台。 2. 单击界面右侧“登录指令”, 获取内网登录指令末尾的SWR镜像仓库地址。例如100.78.15.50:20202。
--log-path	-l	否	日志路径, 不填写时默认为命令行工具当前路径下healthcli.log文件。 路径设置格式: <ul style="list-style-type: none"> <li>Windows系统为“路径\文件名”。</li> <li>Linux系统格式为“路径/文件名”。</li> </ul>
--http-proxy	-p	否	HTTP代理配置, 格式为“http://username:password@your-proxy:your-port”。
--obs-endpoint	-o	否	OBS终端节点名称, 请在 <a href="#">地区与终端节点</a> 中获取。
--obs-install-path	-q	否	设置obsutil安装路径, 默认安装在当前运行目录。 设置时, 该路径必须为obsutil运行文件名, 如/home/path/obsutil、/home/path/obsutil-1.1.1
--obs_down_load_url	-D	否	obsutil下载链接, obsutil将下载到obs-install-path上。 参数有改动时才会触发下载。 下载链接的内容可以是zip、tar.gz文件、二进制文件, 如果是压缩文件, 文件夹内的obsutil必须命名为obsutil (和obsutil官方链接保持一致)。

参数	简写	是否必选	说明
--force	-f	否	强制操作。如果下载obsutil时，指定的obs-install-path上已经有同名文件，不带-f时会提示用户，带上-f会直接覆盖原文件。

● 命令示例

```
health config add -d xxx -u xxx -w xxx -i xxx -r cn-north-4 -o obs.cn-north-4.myhuaweicloud.com -a xxx -s xxx -D https://obs-wwx-2022.obs.cn-north-4.myhuaweicloud.com/obsuti-windows/obsutil.exe -q obsutil-linux/obsutil.exe -f -m xxx -e xxx -t xxx
# 执行成功返回结果如下
add ak successfully!
add sk successfully!
add region successfully!
add platform-id successfully!
add user-name successfully!
add password successfully!
add domain-name successfully!
add obs-endpoint successfully!
add obs_install_path successfully!
add obs_down_load_url successfully!
add iam-endpoint successfully!
add health-endpoint successfully!
add swr-endpoint successfully!
```

图 2-114 命令示例



📖 说明

- 执行以上命令行，会在系统所在的用户目录下自动生成一个.health文件夹，文件夹中包含config.ini配置文件，用于存储任务执行所涉及到的配置，如密钥、区域、当前项目等信息。
- 生成的配置文件不建议直接修改，如需改动请使用命令行工具修改。
- 配置文件中保存有用户的AK、SK信息，为了避免密钥泄露，会对文件中的SK进行加密以保护密钥安全。
- 清空配置请执行health config clear命令。

步骤 4：上传数据

使用upload命令，将本地数据上传到EIHealth平台。该命令不支持将数据上传到引用目录。

最小可以上传0Byte的空文件或文件夹，最大可以上传48.8TB的单个文件。

数据在上传的过程中，受网络影响可能出现损坏，上传命令默认会在上传完成后，验证项目中数据的MD5值与本地数据的MD5值的一致性，以及验证项目中数据的大小与本地数据大小一致性。

- 命令结构  
health upload <srcdir> <destdir> [flags]

表 2-17 参数说明

参数	简写	是否必选	说明
srcdir	无	是	源路径。
destdir	无	是	目的路径。
--rename	-e	否	重命名，上传文件时可选。
--recursive	-r	否	递归上传文件夹中所有的文件和子文件夹，上传文件夹时必选。
--force	-f	否	强制操作，不进行询问提示，上传文件夹时可选。
--flat	-l	否	上传文件夹时，只上传该文件夹下的所有内容，上传文件夹时可选。
--update	-u	否	增量上传操作，设置该参数后上传每个文件时会比对平台数据路径中的文件，仅在以下情况时上传数据： <ul style="list-style-type: none"> <li>文件不存在。</li> <li>待上传文件大小与平台文件大小不一致。</li> <li>文件的最后修改时间不一致。</li> </ul>
--parallel	-p	否	上传文件时，每个分段上传任务数的最大并发数，默认为5。取值范围[1,10]。
--jobs	-j	否	上传文件夹时，批量任务的最大并发数，默认为5。取值范围[1,10]。

- 命令示例  
本节以Windows为例介绍eihealth-toolkit的使用过程，Linux和macOS环境使用方法基本相同，可参考。

- 上传数据前需要使用switch命令进入待操作的项目。例如，使用health switch project test命令进入到名为test的项目中。



```
C:\Users>health switch project test
switch to project test successfully!

C:\Users>
```

- 将本地数据上传到test项目中的src文件夹中。

详细的数据操作命令，如下载、删除、拷贝、切换路径等请参见“[命令行工具 > 数据管理命令](#)”章节。

- 将本地D:\local\data\test.txt路径下的txt文件上传至src文件夹中。  
上传完成后，会在EIHealth平台数据管理页面的src文件中，生成test.txt文件。  

```
health upload D:\local\data\test.txt /src/
```
- 上传文件并重命名为abc.txt。  

```
health upload D:\local\data\test.txt /src/ --rename abc.txt
```
- 递归上传本地D:\local路径下的所有文件和文件夹至项目src文件夹中（不上传local文件夹本身）。  

```
health upload D:\local /src/ -r -f -l
```
- 上传本地D:\local文件夹至项目src文件夹中（上传local文件夹本身）。  

```
health upload D:\local /src/ -r -f
```
- 增量上传文件，同步本地D:\local\test.txt文件至项目src文件夹中。  

```
health upload D:\local\test.txt /src/ --update -r
```

#### 📖 说明

- 列举路径中的对象时，需使用/xxx/格式，如示例中的/src/。
- 列举本地路径中的文件夹对象时，需要使用“路径/文件名”或“路径\文件名”格式。请依据操作系统的路径规范使用，如示例中的D:\local。
- 如果路径中带有特殊字符比如（ ）之类的，运行的时候需要将整个路径用""括起来，或者用\将字符转义，例如：health cp "/src2/abs (1).txt" /src1/ 或者 health cp /src2/abs\(\1).txt /src1/。

## 步骤 5：下载数据

使用**download**命令将EIHealth平台的数据下载到本地，此命令不支持下载引用项目中的数据。

数据在下载的过程中，受网络影响可能出现损坏，下载命令默认会在下载完成后，验证项目中数据的MD5值与本地数据的MD5值的一致性，以及验证项目中数据的大小与本地数据大小一致性。

- 命令结构  

```
health download <srcdir> <destdir> [flags]
```

表 2-18 参数说明

参数	简写	是否必选	描述
srcdir	无	是	源路径。
destdir	无	是	目的路径。
--rename	-e	否	重命名。
--recursive	-r	否	按指定的对象名前缀批量下载，批量下载时必选。
--force	-f	否	强制操作，不进行询问提示，批量下载时可选。

参数	简写	是否必选	描述
--flat	-l	否	批量下载时，不包含上一级父对象名前缀，批量下载时可选。
--update	-u	否	增量下载操作，设置该参数后，下载每个文件时会比对本地的文件，仅在以下情况时下载数据： <ul style="list-style-type: none"> <li>文件不存在。</li> <li>下载文件大小与本地文件大小不一致。</li> <li>文件的最后修改时间不一致。</li> </ul>

- 命令示例

本节以Windows为例介绍eihealth-toolkit的使用过程，Linux和macOS环境使用方法基本相同，可参考。

- a. 下载数据前需要使用switch命令进入待操作的项目。例如，使用health switch project test命令进入到名为test的项目中。
- b. 下载项目中的数据至本地。
  - 下载项目中src/abc.txt文件到本地的data路径中。  
health download /src/abc.txt D:\local\data
  - 下载项目中的abc.txt文件到本地，重命名为abc1.txt。  
health download /src/abc.txt D:\local\data --rename abc1.txt
  - 递归下载项目中src文件夹中的所有文件和文件夹（包含src文件夹本身）至本地data路径，且下载过程中不进行询问提示。  
health download /src/ D:\local\data -r -f
  - 递归下载项目中src文件夹中的所有文件和文件夹（不包含src文件夹本身）至本地data路径，且下载过程中不进行询问提示。  
health download /src/ D:\local\data -r -f --flat
  - 增量下载文件。同步项目中src文件夹下所有的文件和文件夹至本地data路径中。  
health download /src/ D:\local\data --update -r

### 📖 说明

列举路径中的对象时，需使用/xxx/格式，如示例中的/src/。

如果路径中带有特殊字符比如（ ）之类的，运行的时候需要将整个路径用""括起来，或者用\将字符转义，例如：health cp "/src2/abs ( 1 ).txt" /src1/ 或者 health cp /src2/abs\ (1)\.txt" /src1/ 。

## 清理历史命令

为防止配置文件中的敏感信息泄露，建议使用health config clear命令定时清除本地配置文件。

对于执行的历史命令，可通过以下方法清除。

- Linux：执行history -c命令清除历史记录，重新登录命令行工具后，记录可恢复。执行rm -f \$HOME/.bash\_history命令，可删除记录文件，清空历史。

- macOS系统：执行**ps -p \$\$**命令检查shell类型。
  - 类型为bash：操作与Linux系统相同。
  - 类型为zsh：执行**history -p**命令清除历史记录，重新登录命令行工具后，记录可恢复。执行**rm -rf ~/.zsh\_history**命令，退出终端后再次执行**history -p**命令可删除记录文件，清空历史。
- Windows系统：对于当前执行的命令，可通过关闭cmd窗口实现历史命令的清理。

## 2.9 数据库管理

### 2.9.1 数据库简介

EIHealth平台提供了数据库的创建、查询和管理能力。您可以将表型、基因型或其他数据导入EIHealth平台并生成数据库。同时，除了自定义数据库外，支持将作业运行中产生的数据文件创建为数据库。

创建数据库时，数据库模板为您提供了一个数据表的搭建框架，可通过数据库模板灵活的定义数据结构，描述数据及其关系。数据库以项目为粒度进行资源的隔离和访问控制，不同成员角色对数据库的操作权限请参见[成员角色和权限](#)。

### 预置数据库和模板

平台内置“神农项目”药物虚拟筛选数据库和药物配体注释信息数据库，并提供对应的数据库模板。您可以在“数据库”页签单击“shennongProject-database”和“ligandAnnotation-database”数据库查看数据库详情，在“模板”页签单击“shennongProject”和“ligandAnnotation”查看模板详情。

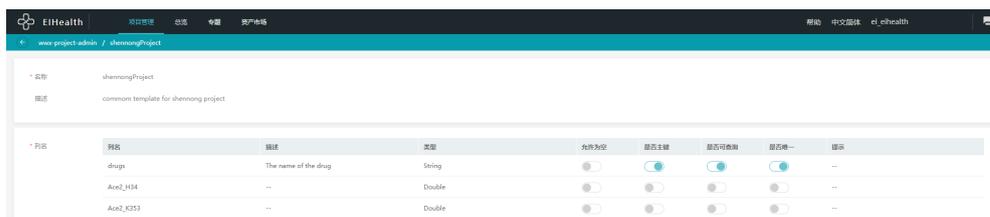
- “神农项目”药物虚拟筛选数据库

数据库呈现了新冠病毒靶点蛋白和药物的结合能信息。“神农项目”是抗疫期间医疗智能体团队联合多家科研单位，从新冠病毒蛋白序列开始，针对所有21个靶点蛋白进行同源建模、分子动力学模拟优化，获取了靶点蛋白的3D结构。并对超过8500个已上市、进入临床的小分子药物进行了约18万种药物-靶点配对情况的计算评估，获取的结合能信息均公开在“神农项目”中。

图 2-115 “神农项目”药物虚拟筛选数据库

drugs 名	Anc2_M34 名	Anc2_K355 名	E 名	Mpro 名	N 名	Nsp1 名	Nsp10 名	Nsp10_site1 名	Nsp10_site2 名	Nsp10_site3 名
5-[3-(2-(4-Cl)phenoxy)propyl]urea	-5.5	-4.2	-5.9	-6.8	-6.6	11.5	-5.8	-6.1	-6.7	-5.8
2-[3-(7-HYDROXY-NALOXONONE-6-yl)propyl]pyridine	-5.1	-3.5	-5.3	-6.8	-6.5	12.9	-7.1	-5.7	-6.3	-5.7
4-Fluorotryptophane	-4.8	-4.3	-4.4	-6	-5.8	-1	-5.7	-5.5	-5.8	-5.5
(5r)-5-Amino-6-Hydroxytryptophan	-4.6	-3.1	-2.7	-4.7	-4.4	-3.1	-4.4	-4	-4.1	-4.1
Similagenin	-5.6	-5.3	-6.9	-7.7	-7.5	48.9	-6.8	-7.5	-8.1	-7.2
BMS-986142	-6.4	-5.5	-7	-8.9	-8.5	62.8	-5.9	-7.8	-7.4	-7.6
3-Methylphenylalanine	-4.3	-4.4	-4.3	-5.2	-5.4	-2.5	-5.4	-4.8	-5.3	-4.8
2-[3-(2-Hydroxy-1,1-dimethylpropyl)amino]propanoic acid	-3.7	-3.3	-2.8	-5.3	-3.9	1.7	-4.4	-4.4	-4.2	-4.3
Bilapratap	-6.3	-5.9	-6.5	-8.2	-7.8	16.1	-6.7	-6.4	-6.9	-6.4
Neobuvir	-4.6	-4.4	-5.3	-7	-6.3	24.3	-5.2	-5.5	-6.3	-5.5

图 2-116 数据库模板



● 药物配体注释信息数据库

数据库呈现了“神农项目”中8500多种药物的注释信息。包含了药物在DrugBank数据库中的ID、药物通用名称、分类、化学式、SMILES结构式、标签和成药性相关的信息。

图 2-117 药物配体注释信息数据库

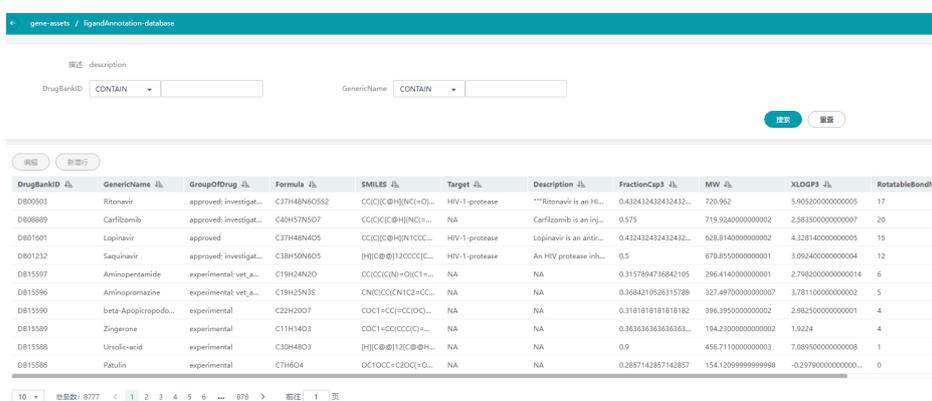
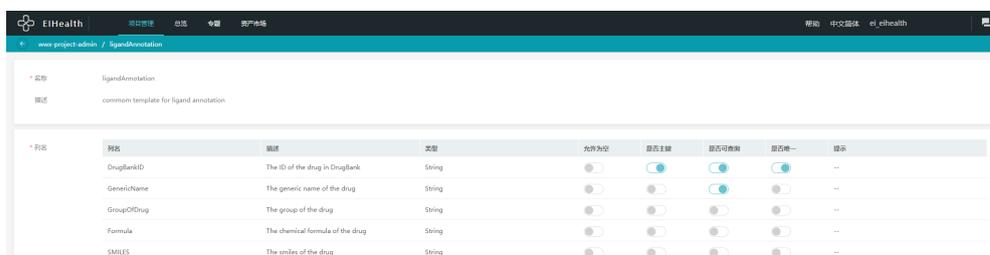


图 2-118 数据库模板



## 2.9.2 创建数据库模板

### 创建数据库模板

创建数据库前，请先在“数据库 > 模板”页签创建数据库模板，模板中数据库表列的数目为固定，行数可按需设置，不同列之间，可通过列信息进行识别。列信息包含列名、描述、类型、允许为空、是否主键、是否可查询、是否唯一、提示和操作信息。单击添加列可新增一列，最多支持100列。单击操作列的🗑️图标可以删除列。

● 列名

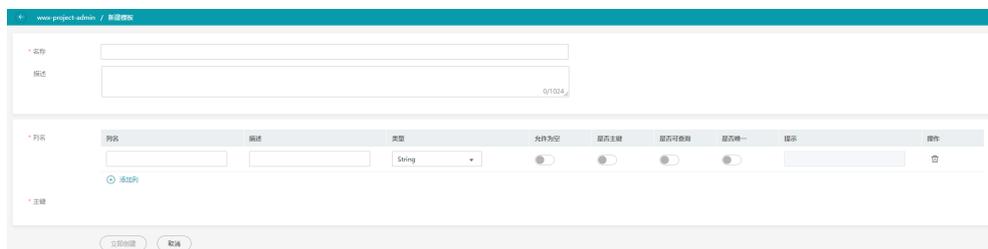
列的名称不可以和PostgreSQL 11版本的**关键字**冲突。

● 是否主键

主键用于唯一地标识数据表中的信息，可通过设置复合主键将数据表中的多个数据设置为唯一，主键至多设置10个。

- 允许为空  
主键所包含的列不允许设置为空。当数据库中某些数据允许没有值时，可设置为空。
- 是否可查询  
依据模板中定义的列信息，给出数据库中可进行查询的数据，可查询至多设置10个。
- 描述  
当参数设置为可查询时，可以添加描述信息。
- 是否唯一  
对数据表中的一个或多个数据设置唯一约束，保证该数据与表中其他行数据不同，是否唯一至多设置10个。

图 2-119 创建数据库模板



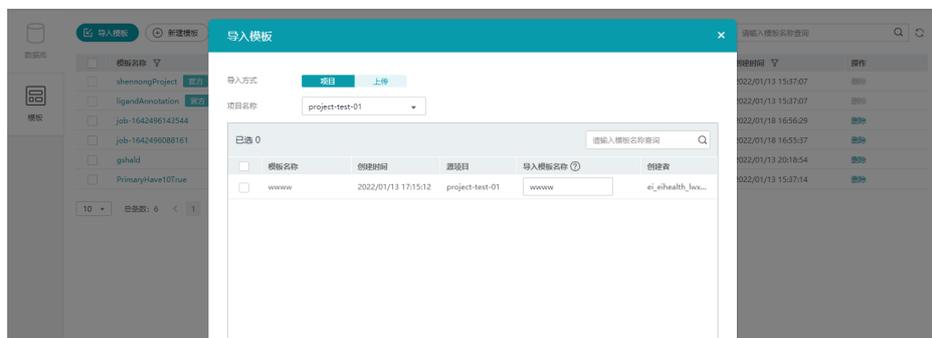
## 2.9.3 导入数据库模板

### 导入模板

提供两种模板导入方式：

- 平台支持导入其他项目的模板，用户需是其他项目中的成员，且为其他项目的“所有者”或“管理员”，预置的模板不支持导入。
- 在左侧导航栏选择“模板”页签，单击“导入模板”。  
导入方式选择“项目”，选择项目名称，勾选需要导入的模板，可以在“导入模板名称”中修改模板名称，单击“确定”。

图 2-120 从其他项目导入模板



- 平台也提供了yaml格式的数据库模板，您可以在本地编辑完成后上传至平台进行使用，数据库模板支持yaml或yml格式，且文件大小不能超过10M。

导入方式选择“上传”，单击“下载示例文件”下载数据库模板示例，编辑后上传模板文件至平台，单击“确定”。

图 2-121 上传模板



模板示例如下：

```

database:
  name: demo # 模板名称, 必填, 长度范围[3,32], 只能包含字母、数字、中划线、下划线, 必须以字母开头
  description: This is demo template # 模板描述, 非必填, 最大长度允许1024
  columns: # 数据库列定义列表, 必填, 最多不能超过100个
    - name: id # 列名, 必填, 长度[1,55], 必须以字母开头, 只能包含字母、数字、下划线
      description: identifier # 列描述, 非必填, 最大长度1024
      type: Long # 列类型, 必填, 支持Long, String, Double三种类型
      nullable: false # 列是否允许未空, 必填
      unique: false # 列是否增加唯一约束, 必填
      primary: true # 列是否作为主键, 必填, 若有多个列设为true, 则多个列组成复合主键
      searchable: true # 列是否可查询, 必填
      tips: uuid # 列的查询提示, 非必填, 若该列为可查询列, 可提供查询提示
    - name: user_name
      description: user name
      type: String
      nullable: false
      unique: false
      primary: false
      searchable: true
      tips: string
    - name: score
      description: score
      type: Double
      nullable: false
      unique: false
      primary: false
      searchable: true
      tips: string
  
```

## 2.9.4 创建数据库

### 创建数据库

数据库支持使用.csv、.txt、.vcf文件生成数据库。创建的数据库需要保证数据文件与模板对应。创建数据库时，可以不选择导入的数据文件，建立空的数据库，后期可以新增数据行或者导入数据。如果使用自动作业的数据表创建数据库，在导入数据，需要参照数据库模板格式进行导入。

您可以导入本项目或使用其他项目中的数据生成数据库。

选择好文件后，通过设置数据的分隔符、跳过的行数来设置有效数据的起始位置。

常见的分隔符有“;”、“,”、“\t”，除此之外您也可以使用其他分隔符。以图 2-122 为例，分隔符为制表符，创建数据库时，分隔符填写为“\t”，有效数据从第 2 行开始，跳过文件行数填写为 1。

### 说明

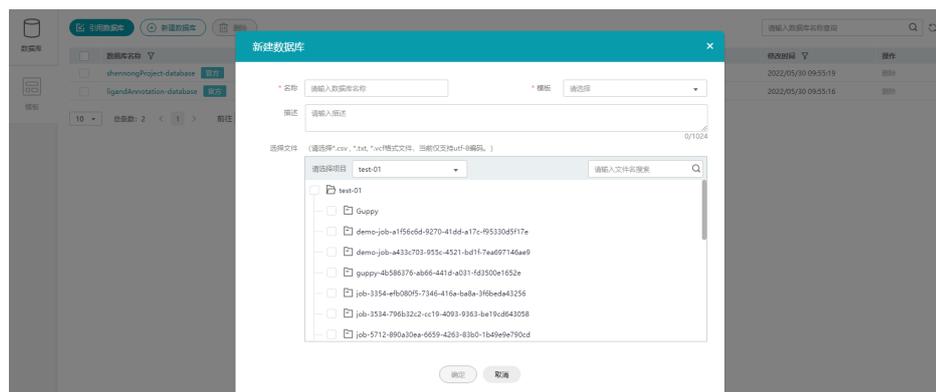
- 在本地编写.csv、.txt、.vcf文件时，文件编码格式使用UTF-8。
- 为保证查询性能，建议数据库行数小于500万行，若超过该量级，建议拆分数据库。
- 导入数据文件大小建议小于1GB。
- 新增的double类型数据，double的取值范围是 $-2^{1024} \sim +2^{1024}$ ，也即 $-1.79E+308 \sim +1.79E+308$ 。超出精度返回会保留小数点后16位的精度。

图 2-122 测试数据

```

USER_NAME → SCORE_LONG → SCORE_DOUBLE
user3 → 10002 → 1.5
user4 → 10000 → 0.4
user5 → 10001 → 0.7
user6 → 10002 → 1.5
user7 → 10002 → 1.5
    
```

图 2-123 创建数据库

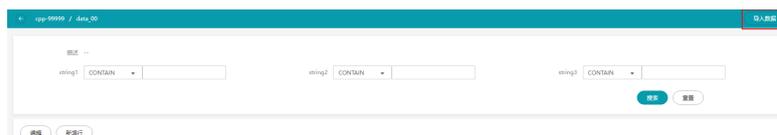


## 导入数据

数据库创建好以后，您可以参照数据库模板格式，将已有的数据进行导入。

1. 单击数据库右上角“导入数据”。

图 2-124 选择导入数据



2. 选择需要导入的数据文件。  
请选择\*.csv, \*.txt, \*.vcf格式文件，当前仅支持utf-8编码。导入数据列与模板保持一致，无需考虑自动作业新建的作业状态更新列。
3. 单击“确定”。

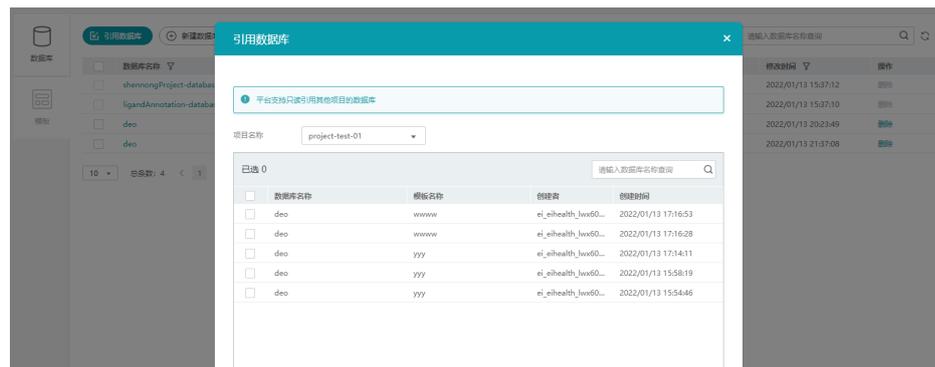
## 2.9.5 引用数据库

### 引用数据库

平台支持引用其他项目的数据库，用户需是其他项目中的成员，且为其他项目的“所有者”或“管理员”。

引用的数据库为只读状态。引用的数据库不支持导入数据。

图 2-125 引用数据库



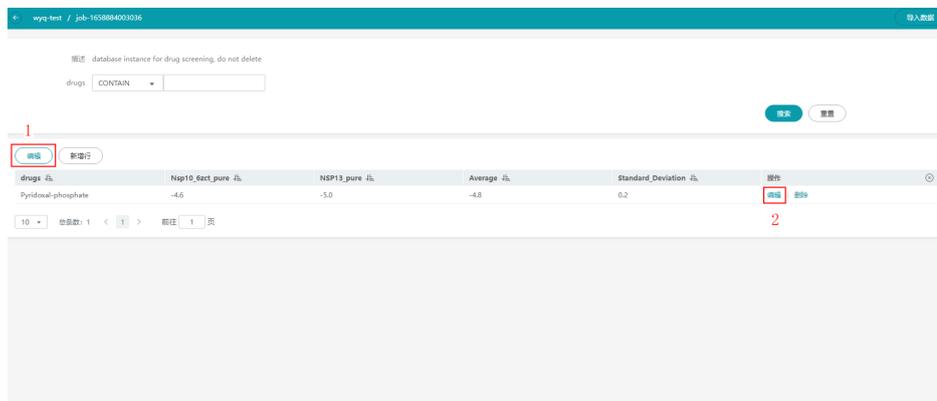
## 2.9.6 管理数据库

### 管理数据库

数据库创建完成后，您可以对数据库内的数据执行编辑、删除、新增行操作。预制数据库和引用数据库不支持编辑、删除、新增行操作。

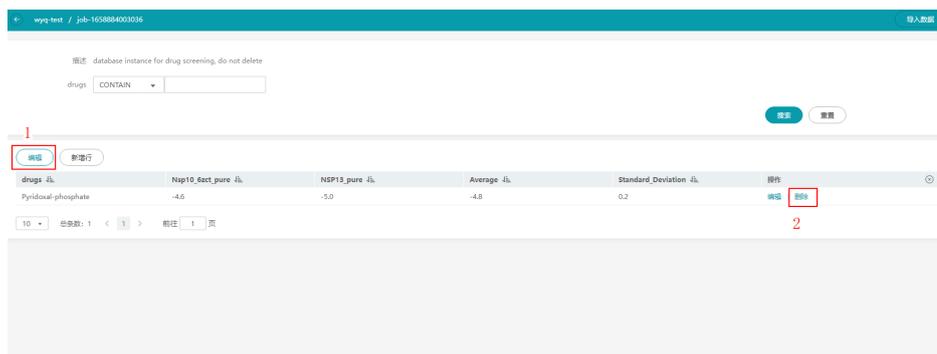
- 编辑
  - a. 数据库创建完成后，单击数据库名称进入数据库详情页，在页面左上角单击“编辑”按钮。
  - b. 在需要修改的数据行的操作列单击“编辑”，修改数据。
    - 编辑时，主键内容不能修改，编辑的值，需要对照数据库模板值进行填写。
    - 对于空的数据库，编辑时，只允许新增行。

图 2-126 编辑数据



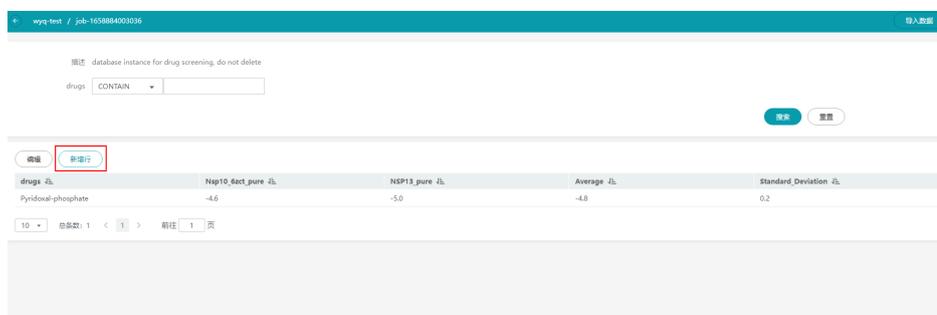
- c. 修改完成后，单击“保存”。
- 删除
  - a. 数据库创建完成后，单击数据库名称进入数据库详情页，在页面左上角单击“编辑”按钮。
  - b. 在需要删除的数据行的操作列单击“删除”。

图 2-127 删除数据



- c. 在删除弹窗中，单击“确定”。
- 新增行
  - a. 数据库创建完成后，单击数据库名称进入数据库详情页，在页面左上角单击“新增行”按钮。

图 2-128 新增行



- b. 在新增的数据行中，填写数据。  
新增的行的数据请按照模板的格式填写。
- c. 填写完成后，单击“保存”。

## 2.10 镜像管理

### 2.10.1 镜像管理简介

运行生物信息学软件，往往由于不同的操作系统（Windows、Linux、Mac等）原因，无法实现统一的运维管理。同时，这些软件具有不同的版本和软件包，安装、使用过程复杂。将生物信息学软件封装成Docker镜像，可以使程序在不同的环境中运行，并通过EIHealth平台的镜像管理，实现高效的调用，极大方便了软件的安装和运行。

Docker镜像是一个模板，是容器应用打包的标准格式，在部署容器化应用时可以指定镜像。例如一个Docker镜像可以包含一个完整的Ubuntu操作系统环境，里面仅安装了用户需要的应用程序及其依赖文件。

EIHealth平台使用容器镜像服务（Software Repository for Container，简称SWR）进行简单易用、安全可靠的镜像管理。单击项目名称，进入所选项目，在“镜像”页签中，以列表形式展示了项目中的镜像。您可查看镜像的详细信息，执行镜像分类、添加描述、删除和查询操作。

图 2-129 镜像管理

名称	源项目	芯片类型	镜像类型	描述	创建时间	更新时间	操作
druglikeness_apitest_k	eihealth_api_project_0	X86	OTHER	--	2022/01/07 15:43...	2022/01/07 15:43...	删除
autodrugglikeness	eihealth_api_project_0	X86	OTHER	--	2022/01/07 15:35...	2022/01/07 15:35...	删除
bwa_apitest_other_loc	eihealth_api_project_0	--	OTHER	--	2022/01/07 15:10...	2022/01/07 15:10...	删除
test_test	eihealth_api_project_0	--	OTHER	--	2022/01/04 17:35...	2022/01/04 17:35...	删除
test-test	eihealth_api_project_0	ARM	OTHER	--	2022/01/04 17:35...	2022/01/04 17:35...	删除
druglikeness	eihealth_api_project_0	--	OTHER	--	2022/01/04 17:25...	2022/01/04 17:25...	删除

## 镜像用途

- 用于创建分析应用

应用是生物信息学软件的镜像封装。例如，您可将Cell Ranger软件封装为镜像，并上传至EIHealth平台。通过应用把镜像引入，利用应用搭建分析流程，执行分析作业。

- 用于创建Notebook

Notebook是一个交互式应用程序，用于代码的编写、调试、运行。创建Notebook时，您可以选择系统镜像。当系统镜像无法满足您的开发需求时，您可以基于EIHealth提供的基础镜像包制作自定义镜像，并上传至平台。您可以在EIHealth平台“开发环境”中使用此自定义镜像创建Notebook。

### 说明

创建Notebook时，如果使用自定义镜像。该自定义镜像，需要基于EIHealth平台提供的基础镜像进行制作。

## 2.10.2 导入或上传镜像

### 镜像导入

镜像按照项目进行划分和管理，隶属于不同项目的镜像可以使用“镜像导入”，导入到本项目中，进行使用。

使用“镜像导入”功能，用户需是其他项目中的成员，且为其他项目的“所有者”或“管理员”。

1. 单击项目名称，进入所选项目，并选择“镜像”，进入镜像管理页面。
2. 单击“镜像导入”。并选择镜像所在的项目、镜像和镜像版本。

图 2-130 镜像导入



- 单击“确定”，完成镜像导入。  
从其他项目导入的镜像，在镜像列表“源项目”列中，显示所属的项目。

## 客户端上传镜像

使用命令行工具eihealth-toolkit上传镜像，详细的上传过程请参见[上传镜像到SWR镜像仓库](#)，详细的命令介绍请参见“[命令行工具 > 镜像管理命令](#)”章节。

### 2.10.3 安装容器引擎

容器引擎是一个开源的引擎，可以轻松地为任何应用创建一个轻量级的、可移植的、自给自足的容器。

容器引擎几乎支持在所有操作系统上安装，用户可以根据需要选择要安装的容器引擎版本。请使用自己的电脑搭建Docker环境，或者使用华为云弹性云服务器ECS搭建Docker环境。

例如，在Linux操作系统下，可以使用如下命令快速安装容器引擎。

```
curl -fsSL get.docker.com -o get-docker.sh
sh get-docker.sh
```

#### 📖 说明

容器镜像服务支持使用容器引擎1.11.2及以上版本上传镜像。

### 2.10.4 获取镜像

#### 获取创建分析应用的镜像

创建分析应用时，您可以通过Docker Hub等镜像仓库，搜索引擎，自己制作等途径获取所需的镜像。

例如，可在Docker Hub获取bwa软件（用于将基因序列比对到参考基因组上）。

#### 📖 说明

以下类型镜像，建议您通过Docker Hub获取，不建议自己制作。

- 基础操作系统类镜像，如Ubuntu、Suse、Centos等。
- 基础编程语言类镜像，如Java、Python、R语言等。
- 基础通用类软件镜像，如Tomcat、Mysql、Ngnix等。

#### 获取创建 Notebook 的镜像

创建Notebook时，平台为您提供了系统镜像和自定义镜像。

- 系统镜像  
工作环境选择PY3版本。
- 自定义镜像  
创建Notebook时所需的自定义镜像，依赖于医疗智能体平台自研的基础镜像，您需要基于获取的基础镜像制作自定义镜像。

先连接容器镜像服务，参考[步骤1.连接容器镜像服务](#)操作，然后使用如下的镜像地址拉取基础镜像。

```
# 基础镜像
docker pull swr.cn-north-4.myhuaweicloud.com/eihealth-notebook/eihealth-notebook:cuda11.0-custom-v1.0.8
```

在基础镜像中，为您内置了CUDA11.0环境。使用CUDA11.0环境前，需要执行以下命令导入环境变量。

```
export PATH=$PATH:/usr/local/nvidia/bin
export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:/usr/local/nvidia/lib64
```

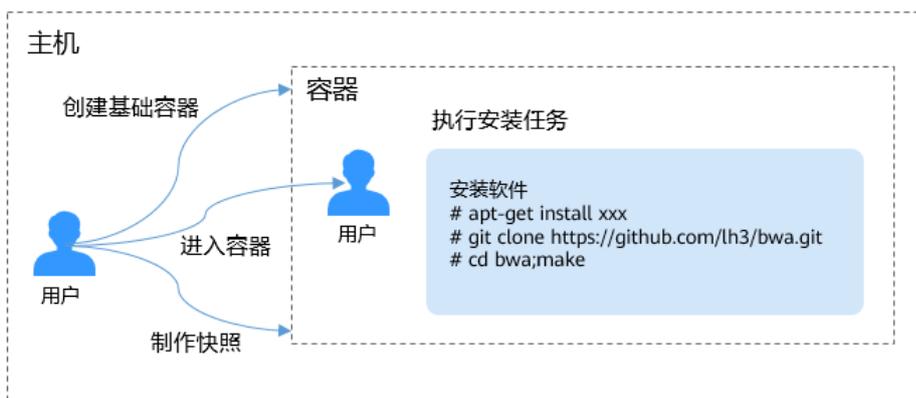
## 2.10.5 制作 Docker 镜像

制作Docker镜像，有以下两种方法。

- **快照方式制作镜像**（偶尔制作的镜像）：在基础镜像上，比如Ubuntu，先登录镜像系统并安装Docker软件，然后整体制作快照，即可得到所需软件的Docker镜像。
- **Dockerfile方式制作镜像**（经常更新的镜像）：将软件安装的流程写成DockerFile，使用Docker build构建成Docker镜像。

### 快照方式制作镜像

如果后续镜像没有变化，可通过快照方式制作镜像。



快照方式制作镜像示例：

本示例中使用华为云[弹性云服务器服务（ECS）](#)创建一台云服务器，并使用快照方式制作**bwa**镜像。

1. 购买弹性云服务器。
2. 云服务器创建成功后，在云服务器列表页，选中待登录的弹性云服务器。单击“远程登录”，输入ECS初始账号，登录ECS。

图 2-131 云服务器列表



3. **安装容器引擎。**
4. 启动一个空白的基础容器，并进入容器。  
例如，启动一个CentOS容器。  

```
docker run -it centos
```
5. 安装依赖包。  

```
yum -y install https://dl.fedoraproject.org/pub/epel/epel-release-latest-8.noarch.rpm  
yum -y install git  
yum -y install gcc automake autoconf libtool make  
yum install -y zlib zlib-devel
```
6. 安装bwa软件，在github上下载bwa的源代码，并使用make编译。  

```
yum install bwa  
git clone https://github.com/lh3/bwa.git  
cd bwa;make
```

### 📖 说明

请预先安装好Git，并检查本机是否有ssh key设置。

7. 输入**exit**退出容器。
8. 查询容器id。  

```
docker ps -a
```
9. 制作快照。  

```
docker commit -m "xx" -a "tsj" container-id tsj/image:tag
```

  
例如：

```
docker commit -m "test" -a "username" adb1127979a1 bwa:v0.7
```

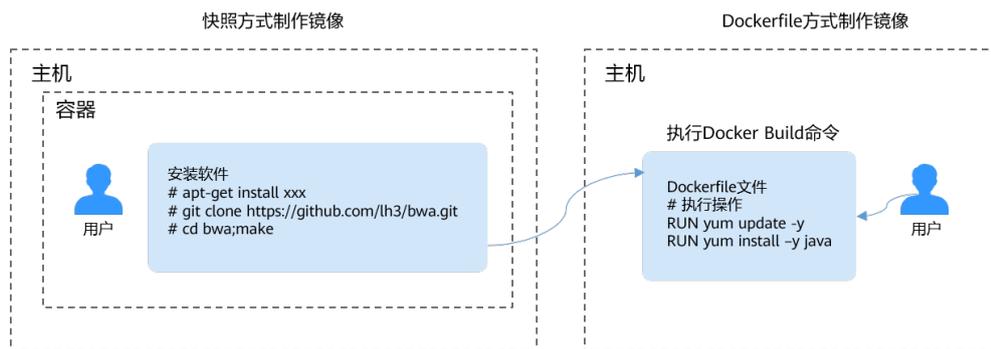
  - -a: 提交的镜像作者，例如tsj。
  - container-id: 容器id。
  - -m: 提交时的说明文字，例如xx。
  - tsj/image:tag: 仓库名/镜像名:TAG名，名称可自定义。
10. 执行**docker images**命令查看制作完成的Docker镜像。

## Dockerfile 方式制作镜像

如果后续镜像经常变更（例如某个软件更新版本），建议使用Dockerfile方式制作镜像。如果采用快照方式制作镜像，则每次变更都需要执行操作命令，制作过程较为繁琐。

Dockerfile方式制作镜像是将快照制作的方式用Dockerfile文件写出来，然后执行**docker build -t tsj/image:tag .**命令，自动完成镜像制作。

命令中“.”表示DockerFile文件的路径，“tsj/image:tag”表示仓库名/镜像名:TAG名。



Dockerfile方式制作镜像常用命令示例：

本例中使用华为云弹性云服务器服务（ECS）创建一台云服务器，制作centos镜像，并在镜像中放入图片文件，讲解Dockerfile常用命令使用方法。

1. [购买弹性云服务器](#)。
2. 云服务器创建成功后，在云服务器列表页，选中待登录的弹性云服务器。单击“远程登录”，输入ECS初始账号，登录ECS。

图 2-132 云服务器列表



3. [安装容器引擎](#)。
4. 创建一个名为workdir的目录。  
mkdir workdir
5. 在该目录中创建一个Dockerfile文件。  
cd ./workdir  
touch Dockerfile
6. 在该目录中创建一个空白文件abc.txt，创建一个webapp目录。  
touch abc.txt  
mkdir webapp
7. 在该目录中下载一个网络图片。  
wget https://www.baidu.com/img/bd\_logo1.png
8. 执行vi Dockerfile命令，进入Dockerfile文件中，编写文件。

```
#Version 1.0.1
FROM centos:latest
MAINTAINER ***u "***u@xxx.com"

#设置root用户为后续命令的执行者
USER root

#执行操作
RUN yum update -y
RUN yum install -y java

#使用&&拼接命令
RUN touch test.txt && echo "abc" >>abc.txt

#对外暴露端口
EXPOSE 80 8080 1038

#复制文件
COPY abc.txt /opt/
```

```
#复制文件夹
COPY /webapp /opt/webapp

#复制图片文件
COPY bd_logo1.png /opt/

#设置环境变量
ENV WEBAPP_PORT=9090

#设置工作目录
WORKDIR /opt/

#设置启动命令
ENTRYPOINT ["ls"]

#设置启动参数
CMD ["-a", "-l"]

#设置卷
VOLUME ["/data", "/var/www"]

#设置子镜像的触发操作，该命令在创建镜像时不会执行，只有在后续依照该镜像创建新镜像时才会执行
ONBUILD ADD . /app/src
ONBUILD RUN echo "on build excuted" >> onbuild.txt
```

9. 按键盘Esc键，并执行:wq保存退出Dockerfile。

10. 制作镜像。

```
docker build -t centos:v01 .
```

详细的Dockerfile指令请参见[Dockerfile参考](#)、《[容器镜像服务 编写高效的Dockerfile](#)》。

### 📖 说明

使用获取的Notebook基础镜像制作自定义镜像时，请参考以下说明。

- Dockerfile中不可以指定CMD以及ENTRYPOINT，否则会覆盖基础镜像启动脚本，引起异常。
- 基础镜像中的PyPi Mirror，默认配置为华为云软件开发云的PyPi mirror。您可以在容器中执行如下命令，查看PyPi Mirror。如果您想用其他PyPi Mirror，可将命令中的index-url参数修改为您需要的PyPi mirror。

```
cat /root/.pip/pip.conf.product
[global]
index-url = http://repo.myhuaweicloud.com/repository/pypi/simple
format = columns
[install]
trusted-host=repo.myhuaweicloud.com
```

## Dockerfile 基本语法

- FROM  
指定待扩展的父级镜像(基础镜像)。除了注释以外，在文件的开头必须是一个FROM指令，后面的指令便在这个父级镜像的环境中运行，直到遇到下一个FROM指令。通过添加多个FROM命令，可以在同一个Dockerefile文件中创建多个镜像。
- MAINTAINER  
声明创建镜像的作者信息：用户名、邮箱，非必须参数。
- RUN  
用来修改镜像的命令，常用来安装库、程序以及配置程序。一条RUN指令执行完毕后，会在当前镜像上创建一个新的镜像层，接下来的指令会在新的镜像上继续执行。

RUN 语句具有以下形式。

- **RUN yum update**: 在/bin/sh路径中执行的指令命令。
- **RUN ["yum", "update"]**: 直接使用系统调用exec来执行。
- **RUN yum update && yum install nginx**: 使用&&符号将多条命令连接在同一条RUN语句中。

- EXPOSE

用来指明容器内进程对外开放的端口，多个端口之间使用空格隔开。运行容器时，通过参数-P(大写)即可将EXPOSE里所指定的端口映射到主机上另外的随机端口，其他容器或主机就可以通过映射后的端口与此容器通信。您也可以通过-p(小写)参数将Dockerfile中EXPOSE中没有列出的端口设置成公开的。

- COPY

将本地的文件或目录复制到镜像中。

- ADD

向新镜像中添加文件，这个文件可以是主机文件、网络文件或文件夹。

- 第一个参数：源文件（夹）。
  - 如果是相对路径，必须是相对于Dockerfile所在目录的相对路径。
  - 如果是URL，会将文件先下载下来，然后再添加到镜像里。
- 第二个参数：目标路径。
  - 如果源文件是主机上的zip或者tar形式的压缩文件，Docker会先解压缩，然后将文件添加到镜像的指定位置。
  - 如果源文件是一个通过URL指定的网络压缩文件，则不会解压。

- VOLUME

在镜像里创建一个指定路径(文件或文件夹)的挂载点，这个容器可以来自主机或者其它容器。多个容器可以通过同一个挂载点共享数据，即便其中一个容器已经停止，挂载点也仍然可以访问。

- WORKDIR

为接下来执行的指令指定一个新的工作目录，这个目录可以是绝对目录，也可以是相对目录。根据需要，WORKDIR可以被多次指定。当启动一个容器时，最后一条WORKDIR指令所指的目录将作为容器运行的当前工作目录。

- ENV

设置容器运行的环境变量。在运行容器的时候，通过设置-e参数可以修改这个环境变量值，也可以添加新的环境变量。

例如：

```
docker run -e WEBAPP_PORT=8000 -e WEBAPP_HOST=www.example.com ...
```

- CMD

用来设置启动容器时默认运行的命令。

- ENTRYPOINT

用来指定容器启动时的默认运行的命令，与CMD类似。区别在于：运行容器时添加在镜像之后的参数，对ENTRYPOINT是拼接，CMD是覆盖。

- 若在DockerFile中指定了容器启动时的默认运行命令为**ls -l**，则运行容器时默认启动命令为**ls -l**。

- **ENTRYPOINT [ "ls", "-l" ]**: 指定容器启动时的程序及参数为ls -l。
  - **docker run centos**: 当运行centos容器时，默认执行的命令是**docker run centos ls -l**。
  - **docker run centos -a**: 当运行centos容器时拼接了-a参数，则默认运行的命令是**docker run centos ls -l -a**。
- 若在DockerFile中指定了容器启动时的默认运行命令为**--entrypoint**，则在运行容器时若需要替换默认运行命令，可以通过添加**--entrypoint**参数来替换Dockerfile中的指定。
- ```
docker run gutianlangyu/test --entrypoint echo "hello world"
```
- **USER**  
为容器的运行及RUN、CMD、ENTRYPOINT等指令的运行指定用户或UID。
  - **ONBUILD**  
触发器指令。构建镜像时，Docker的镜像构建器会将所有的ONBUILD指令指定的命令保存到镜像的元数据中，这些命令在当前镜像的构建过程中并不会执行。只有新的镜像使用FROM指令指定父镜像为当前镜像时，才会触发执行。  
使用FROM以这个Dockerfile构建出的镜像为父镜像，构建子镜像时：  
**ONBUILD ADD . /app/src**: 自动执行**ADD . /app/src**

## 2.10.6 上传镜像到 SWR 镜像仓库

1. 上传镜像前，您需要安装容器引擎并完成镜像的制作，详细操作请参考[安装容器引擎、制作Docker镜像](#)。如果您已经安装了容器引擎，请跳过该步骤。
2. [下载命令行工具](#)并进行[初始化配置](#)。
3. 使用**health switch project**命令进入到所需的项目中。  
# 命令结构  
health switch project <project-name>  
# 命令示例，例如进入到名为demo-project的项目中  
health switch project demo-project
4. 上传镜像到SWR镜像仓库。
  - a. 使用**health docker tag**命令给要上传的镜像打标签。  
source-image-name为待上传镜像名称，source-tag-name为待上传镜像标签，target-image-name和target-tag-name可自定义。  
# 命令结构  
health docker tag <source-image-name:source-tag-name> <target-image-name:target-tag-name>  
# 命令示例  
health docker tag demo-image:v1.0 project-demo-image:v1.0
  - b. 使用**health docker push**命令上传镜像。  
# 命令结构  
health docker push <image-name:tag-name> [flags]  
# 命令示例  
health docker push project-demo-image:v1.0

执行**health docker images**命令查看已有的镜像。  
详细的命令介绍请参见 [“命令行工具 > 镜像管理命令”](#) 章节。
5. 单击“镜像”，在镜像列表中查看已上传的镜像。

图 2-133 镜像列表

| 名称                      | 源项目                    | 芯片类型 | 镜像类型  | 描述 | 创建时间                 | 更新时间                 | 操作 |
|-------------------------|------------------------|------|-------|----|----------------------|----------------------|----|
| druglikeness_apitest_1c | eihealth_api_project_0 | X86  | OTHER | -- | 2022/01/07 15:43:... | 2022/01/07 15:43:... | 删除 |
| autodruglikeness        | eihealth_api_project_0 | X86  | OTHER | -- | 2022/01/07 15:35:... | 2022/01/07 15:35:... | 删除 |
| bwa_apitest_other_loc   | eihealth_api_project_0 | --   | OTHER | -- | 2022/01/07 15:10:... | 2022/01/07 15:10:... | 删除 |
| test_test               | eihealth_api_project_0 | --   | OTHER | -- | 2022/01/04 17:35:... | 2022/01/04 17:35:... | 删除 |
| test-test               | eihealth_api_project_0 | ARM  | OTHER | -- | 2022/01/04 17:35:... | 2022/01/04 17:35:... | 删除 |
| druglikeness            | eihealth_api_project_0 | --   | OTHER | -- | 2022/01/04 17:25:... | 2022/01/04 17:25:... | 删除 |

6. 【可选】单击“镜像类型”，对上传的镜像进行分类。  
上传镜像时如果上传命令中未指定镜像类型，新上传的镜像默认显示为“OTHER”，您可以将镜像标记为“APP”或“NOTEBOOK”。
- APP：用于创建应用的镜像。
  - NOTEBOOK：用于创建Notebook的镜像。

### 须知

为了保证平台业务安全，您在平台内购买的计算资源，将部署在独立的、专属资源池中。Notebook与流程作业均在此资源池运行，其中Notebook启动用户为health-user，流程作业默认启动用户为root。

平台提供了必要的加固措施以保证容器安全、隔离性及数据安全，但由于容器技术本身的限制，仍可能存在容器逃逸等问题，进而导致越权访问。

为避免此类问题发生，请您确保上传镜像来源的合法性、有效性，禁止上传未知来源、恶意程序等非法镜像。

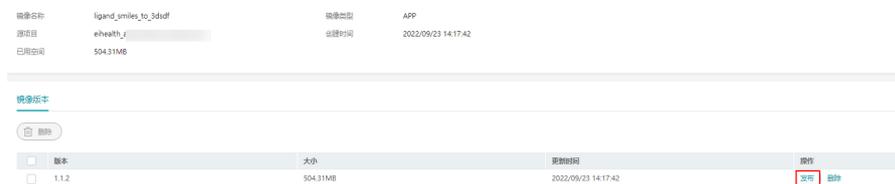
## 2.10.7 发布镜像

您可以将创建的镜像发布到资产市场，供其他项目订阅和使用。

发布资产前，需要先设置商标。详细可参考[商标设置](#)。

1. 单击项目名称，进入项目管理页面，选择“镜像”页签。
2. 单击需要发布的镜像名称，进入镜像详情页面。
3. 单击“操作”列的“发布”。

图 2-134 发布镜像



4. 在发布页面，填写发布镜像的相关信息。

表 2-19 发布镜像参数说明

| 参数   | 说明                                                                                                                                  |
|------|-------------------------------------------------------------------------------------------------------------------------------------|
| 名称   | 发布镜像的名称。名称长度为1-128个字符，以小写字母或数字开头，只允许出现中划线(-)、下划线(_)、小写字母和数字，且必须以小写字母或数字结尾。                                                          |
| 展示名称 | 发布成功后在组织共享中展示的名称。若不填写，展示的名称和镜像名称相同。发布成功的镜像不支持修改展示名称。                                                                                |
| 版本   | 发布镜像的版本。<br>版本长度必须为1-64个字符，只支持字母，数字，中划线，下划线和点号。                                                                                     |
| 类型   | 镜像。                                                                                                                                 |
| 标签   | 设置发布镜像的标签。<br>最多支持设置5个标签。                                                                                                           |
| 短描述  | 为发布的镜像设置的简要描述。                                                                                                                      |
| 图片   | 展示在资产市场的图片。要求如下： <ul style="list-style-type: none"> <li>● 格式：svg,png,jpg,jpeg</li> <li>● 尺寸：16 * 9</li> <li>● 大小：不超过20KB</li> </ul> |
| 描述   | 针对此镜像的相关描述。                                                                                                                         |

5. 单击“立即发布”。

系统跳转到“我的发布”页面，可以查看发布的镜像状态。当镜像发布成功后，可以在“资产市场>组织共享”中查看发布的镜像。

 说明

- 已发布的镜像不能直接编辑和删除，需下线后才能编辑和删除。
- 发布中的镜像可以取消，取消后状态为发布失败。
- 已发布的镜像可以订阅和下线。
- 同一个发布者可以发布同名称，同类型的镜像。若版本不同，则合并。用户在使用时，可以选择对应的版本号进行订阅或者下线。若版本相同，则更改版本后发布。
- 已发布的镜像，可以在“资产市场>我的发布”中进行下线。下线后已经订阅的镜像不可以使用。

## 2.11 工具管理

### 2.11.1 工具管理简介

EIHealth中的每一个分析作业都依托于应用运行。应用可以组合形成分析流程。

应用是生物信息学软件的镜像封装。您可以将软件制作成镜像，并将镜像上传至EIHealth平台，通过应用引入镜像。制作好的应用可以单独使用，也可以将多个应用

编排入流程串联使用。在“项目管理”页面“工具”页签中，以列表形式展示了项目中的应用。您可以[新建应用](#)、[导入应用](#)或[上传应用](#)，并查看应用详情、版本、创建者、修改和创建时间，可以对名称、创建者、修改时间、创建时间、源项目进行排序。并可执行查询、修改和删除应用的操作。

图 2-135 应用列表

| 名称                    | 图标 | 源项目                              | 版本    | 创建者                    | 创建时间                | 修改时间                | 操作           |
|-----------------------|----|----------------------------------|-------|------------------------|---------------------|---------------------|--------------|
| xpp-000               |    | cpd-99999                        | 1.0   | et_ehealth_hen60784... | 2022/09/27 15:55:15 | 2022/09/27 15:55:15 | 编辑   删除   更多 |
| ligand_mimes_to_3doff |    | molecular docking drug discovery | 1.0.0 | et_ehealth_hen60784... | 2022/09/27 15:45:12 | 2022/09/27 15:45:12 | 编辑   删除   更多 |

流程包含分析过程中所需应用的执行信息和数据的输入、输出等参数定义。流程通过流程设计器创建，创建好的流程将存储于“项目管理”页面“工具”页签中。在该页签中，以列表形式展示了项目中的流程。您可以[新建流程](#)、[导入流程](#)或[上传流程](#)，并查看流程详情、版本、创建者、修改和创建时间，可以对名称、创建者、修改时间、创建时间、源项目进行排序。并可执行查询、编辑修改流程、删除操作和基于该流程创建分析作业。

图 2-136 流程列表

| 名称      | 图标 | 源项目                 | 版本    | 创建者                    | 创建时间                | 修改时间                | 操作           |
|---------|----|---------------------|-------|------------------------|---------------------|---------------------|--------------|
| RNA-seq |    | Genome RNA-Seq STAR | 1.0.0 | et_ehealth_hen60784... | 2022/09/27 17:11:07 | 2022/09/27 17:11:07 | 编辑   删除   更多 |
| xpp-000 |    | cpd-99999           | 1.0   | et_ehealth_hen60784... | 2022/09/27 15:55:28 | 2022/09/27 15:55:28 | 编辑   删除   更多 |

## 2.11.2 新建应用

详细的创建应用过程、镜像填写方法、参数填写方法，请参考[创建应用样例](#)章节。

1. 单击“新建应用”，进入新建应用页面。

图 2-137 新建应用

2. 填写应用的基本信息，包括“名称”、“版本”、“图标”、“标签”、“短描述”和“描述”。

图 2-138 基本信息

\* 名称

版本

图标

标签

短描述

描述

最近常用标签: 宏观脑科学 | 脑科学 | 功能磁共振

3. 选择镜像和镜像版本。

详细的镜像介绍和制作方法请参见[镜像管理](#)。

4. 填写镜像启动命令。

镜像启动命令需要引用输入、输出参数中的变量，并以大括号扩起，以\$符号进行引用。

镜像启动命令支持多行输入，每行最多256字符，最多支持300行。

例如，bamqc软件输入参数填写为genomedir、thread和input，输出参数为outputdir，则镜像启动命令如下所示。

```
bamqc -t ${thread} -g ${genomedir} --noextract -o ${outputdir} ${input}
```

图 2-139 镜像信息

The screenshot shows two form fields. The first is labeled '\* 选择镜像' (Select Image) and contains a text input field with the placeholder '请选择容器镜像' (Please select container image) and a '选择镜像' (Select Image) button. The second is labeled '\* 镜像启动命令' (Image Start Command) and contains a large text area with the placeholder '请输入镜像启动命令' (Please enter image start command).

5. 选择CPU、GPU类型和大小，选择内存大小，内存单位为GB。

- CPU架构依赖于制作镜像过程中选择的系统类型，以及制作镜像时所需的生物信息学软件支持在X86还是ARM上运行。例如，GATK是基于X86指令集开发的生信软件，使用CentOS的X86系统创建GATK镜像，则在创建应用时选择“X86”。
- CPU需求：请按实际需求填写，取值范围为“0.1-128”，单位C，支持一位小数，不填默认1C。
- Memory需求：请按实际需求填写，取值范围为“0.1-3072”，单位GB，支持一位小数，不填默认1GB。
- GPU类型：请按实际需求填写，取值范围为“无、GPU、D310、Snt9”，如果选择Snt9，GPU需求需要是0、1、2、4、8。
- GPU需求：请按实际需求填写，只能输入0到16的正整数。
- 计算节点标签：请选择标签名称，不支持多选。应用将会调度到有相应节点标签的计算节点。计算节点标签设置方法请参见[计算资源标签管理](#)。

6. 填写输入参数、输出参数。

参数填写时，输入参数及输出参数有字符串（String），文件（File），文件夹（Directory），枚举（Enum）四种类型。

图 2-140 参数信息

The screenshot shows a table for configuring parameters. The table has columns for parameter name, data type, parameter mode, description, required, parallel, other, and operation. There are two rows of parameter configuration visible, one for input parameters and one for output parameters. Each row includes a text input field for the parameter name, a dropdown for data type (File is selected), a text input for the parameter mode, a text input for the description, a toggle for required, a toggle for parallel, a text input for other, and a delete icon.

## 说明

- 对于输入参数，打开“并发”开关，在启动作业时，每个参数可以设置多个参数值，自动生成多个作业并发执行。并发执行的作业数为设置的参数值个数的乘积。  
例如，存在输入参数a和输入参数b，在启动作业时，分别给参数a设置了2个参数值，给参数b设置了2个参数值。那么，系统将自动生成4个作业并发执行。
  - 对于输出参数，如果镜像启动命令中指定了输出参数，则在设置输出参数时，需要勾选“必传”，并填写“默认值”。
    - 如果输出参数为Directory、File类型，默认值必须以/开头，且不能以/结尾，开头的/代表本项目的桶路径。
    - 如果输出参数为String，默认值最大长度不超过256，仅支持字母、数字、中划线、下划线、小数点和斜线。
    - 如果输出参数为Enum类型，需填写有效值，并在有效值中选择默认值。
7. 单击“立即创建”，创建应用。

## 2.11.3 导入应用

导入应用是将隶属于其他项目的应用导入至本项目中，一次至多导入50个应用。

使用“导入应用”功能，用户需是其他项目中的成员，且为其他项目的“所有者”或“管理员”。

- 单击“导入应用”，进入导入应用页面。

图 2-141 导入应用



- 选择需要引用的项目以及项目中的应用，选择应用的版本。“导入应用名称”是选填项，可以使用原有名称，或自定义。

图 2-142 导入应用



- 单击“确定”，导入应用，导入过程中可以查看每个应用的导入状态。

## 2.11.4 上传应用

通过在本地修改应用的yaml模板，上传应用至项目中。

- 获取应用yaml模板。

- 单击“上传应用”，在弹出的页面中下载yaml示例文件。
- 使用命令行工具，执行**health get app -s**命令获取创建应用的yaml模板，复制模板并保存到本地。可以保存成.yaml或.txt文件，保存为txt文件时，其内容需为yaml格式。

## 2. 修改应用yaml模板。

在本地编写模板，模板中的参数与在EIHealth平台**新建应用**过程一致。

修改yaml模板时，详细的命令和参数请参见[应用配置文件说明](#)。

```
# 详情说明可参考API文档中应用管理-创建应用
app:
  id: "                # 应用id
  name: 'demo-app'    # 应用名称 取值范围: [1,56], 以大小写字母开头, 允许出现中划线(-)、下划线(_)、大小写字母和数字, 且必须以大小写字母或数字结尾。更新应用时, 应用名称不支持修改。
  version: '1.0.0'    # 应用版本 取值范围: [1,24], 以小写字母或数字或大写字母开头, 允许出现中划线, 必须以小写字母或数字或大写字母结尾。更新应用时, 应用版本不支持修改。
  summary: 'summary'  # 应用简述 取值范围[0,128]
  description: 'description' # 应用描述 取值范围[0,65535], 后续支持markdown文本
  labels:             # 应用标签 取值范围[0,5], 标签名称长度为1~32, 支持中文、字母、数字、空格、下划线和中划线, 且不能以空格开头或者结尾。
  - labelA
  - labelB
  image: 'gwj-test-01/busybox:latest' # docker镜像地址取值范围[1-255], 不能包含中文字符, 格式为{project-name}/{image-repo}:{image-tag}
  commands:           # docker启动时执行命令 单行命令长度取值范围[0-256], 不能包含中文字符, 支持多行输入, 最多支持300行。
  - 'echo eihealth;'
  node_labels:        # 计算节点标签 标签个数取值范围[0-1],单个节点标签最大长度56个字符, 以字母或数字开头, 可以包含下划线(_)、中划线(-)、点(.)、字母和数字, 且必须以字母或数字结尾
  - health.label
  resources:          # 应用的资源配额
  cpu_type: 'X86'     # cpu架构类型, 不填默认X86
  cpu: '0.1C'        # cpu申请使用量, 取值范围[0.1-128], 单位C, 支持一位小数。对于应用, 不填默认1C; 对于流程和作业, 不填默认使用前一级的配置, 填值会覆盖更新。覆盖关系: 作业->流程->应用
  memory: '0.1G'     # 内存申请使用量, 取值范围[0.1-3072], 单位G, 支持一位小数。对于应用, 不填默认1G; 对于流程和作业, 不填默认使用前一级的配置, 填值会覆盖更新。覆盖关系: 作业->流程->应用
  gpu_type: "        # gpu架构类型, 取值范围 ' |GPU|Snt9|D310。对于流程和作业, 不填默认使用前一级的配置, 填值会覆盖更新。覆盖关系: 作业->流程->应用
  gpu: '0'           # gpu申请使用量, 取值范围[0-16], 仅支持整数, Snt9有特殊约束, 申请数量需要是0,1,2,4,8。
                    # 对于应用, 不填默认0; 对于流程和作业, 不填默认使用前一级的配置, 填值会覆盖更新。覆盖关系: 作业->流程->应用
  inputs:
  - name: 'input-dir' # 参数名称, 单个应用内唯一。取值范围: 长度为[1,32], 以小写字母开头, 允许出现中划线(-)、小写字母和数字, 且必须以小写字母或数字结尾。
    type: DIRECTORY  # 参数类型。取值: [STRING, FILE, DIRECTORY, ENUM]
    pattern: '*.fastq' # 提示用户参数填写的格式, 取值范围: [0-64]。对于STRING类型, 匹配字符串内容, 比如后缀约束.fastq;
                    # 对于ENUM类型, 可以提示一定要在param_enum列表范围内取值; 对于FILE类型, 约束文件后缀类型; 对于DIRECTORY类型, 提示xxx;
    required: true    # 参数是否必须, 取值[true,false]
    concurrent: vars_iter # 是否并发, 取值 vars_iter 时为开启并发状态, 不设置则不开启
    description: "    # 参数描述。取值范围: [0-255]
    values:          # 参数取值 如填写, 只支持填一项, 根据参数类型进行不同的校验
  - 'gwj-test-01:/test'
  - name: 'input-enum' # 参数名称, 单个应用内唯一。取值范围: 长度为[1,32], 以小写字母开头, 允许出现中划线(-)、小写字母和数字, 且必须以小写字母或数字结尾。
    type: ENUM        # 参数类型。取值: [STRING, FILE, DIRECTORY, ENUM]
    pattern: "        # 提示用户参数填写的格式, 取值范围: [0-64]。对于STRING类型, 匹配字符串内容, 比如后缀约束.fastq;
                    # 对于ENUM类型, 可以提示一定要在param_enum列表范围内取值; 对于FILE类型, 约束文件后缀类型; 对于DIRECTORY类型, 提示xxx;
    required: true    # 参数是否必须, 取值[true,false]
    description: "    # 参数描述。取值范围: [0-255]
    enum:            # type 为 ENUM 类型时需要添加enum选项
```

```
- test
- test2
values: # 参数取值 如填写，只支持填一项，只支持enum内的选项
- test
outputs:
- name: 'output-dir'
  type: DIRECTORY
  pattern: '*'
  required: true
  description: "
  values:
  - '/output'
```

3. 上传应用模板。  
单击“上传应用”，在弹出的页面中上传yaml模板。

## 2.11.5 发布应用

您可以将创建的应用发布到资产市场，供其他项目订阅和使用。

发布资产前，需要先设置商标。详细可参考[商标设置](#)。

1. 在“工具”页面选择“应用”。
2. 选择需要发布的应用，单击“操作”列“更多”>“发布”。
3. 在发布页面，填写发布应用的相关信息。

表 2-20 发布应用参数说明

| 参数   | 说明                                                                                                                                  |
|------|-------------------------------------------------------------------------------------------------------------------------------------|
| 名称   | 发布应用的名称。名称长度为1-56，以字母开头，只允许出现中划线(-)、下划线(_)、字母和数字，且必须以字母或数字结尾。                                                                       |
| 展示名称 | 发布成功后在组织共享中展示的名称。若不填写，展示的名称和应用名称相同。发布成功的应用不支持修改展示名称。                                                                                |
| 版本   | 发布应用的版本。<br>版本长度必须为1-24个字符，以字母或数字开头（多个字符可以连续），中间只能包含-特殊符号（多个特殊符号不能连续），以字母或数字结尾。                                                     |
| 类型   | 应用。                                                                                                                                 |
| 标签   | 设置发布应用的标签。<br>最多支持设置5个标签。                                                                                                           |
| 短描述  | 为发布的应用设置的简要描述。                                                                                                                      |
| 图片   | 展示在资产市场的图片。要求如下： <ul style="list-style-type: none"> <li>● 格式：svg,png,jpg,jpeg</li> <li>● 尺寸：16 * 9</li> <li>● 大小：不超过20KB</li> </ul> |
| 描述   | 针对此应用的相关描述。                                                                                                                         |

4. 单击“立即发布”。

系统跳转到“我的发布”页面，可以查看发布的应用状态。当应用发布成功后，可以在“资产市场>组织共享”中查看发布的应用。

**说明**

- 已发布的应用不能直接编辑和删除，需下线后才能编辑和删除。
- 发布中的应用可以取消，取消后状态为发布失败。
- 已发布的应用可以订阅和下线。
- 同一个发布者可以发布同名称，同类型的应用。若版本不同，则合并。用户在使用时，可以选择对应的版本号进行订阅或者下线。若版本相同，则更改版本后发布。
- 已发布的应用，可以在“资产市场>我的发布”中进行下线。下线后已经订阅的应用不可以使用。

## 2.11.6 编辑应用

您可以对已经创建的应用进行编辑，修改对应参数信息。

1. 在应用列表页面，选择需要修改的应用，单击操作列“编辑”。

图 2-143 编辑应用



2. 在编辑应用页面，可以修改相关的参数。相关参数描述可参考[新建应用](#)。

**说明**

如果更改了版本号，相当于新建一个应用。

## 2.11.7 创建 FastQC 应用样例

本章节提供了在EIHealth平台创建FastQC应用的样例，帮助您快速熟悉平台的使用方法。

FastQC是一款高通量序列数据的质量检测工具，此样例基于开源的FastQC软件，将软件制作成镜像，上传至平台，并基于此镜像创建应用。应用创建完成后可以直接使用FastQC应用，或将其编排至流程，和其他应用一起使用。

创建FastQC应用时，需要您熟悉该软件的使用，并具备一定的开发经验。FastQC的详细介绍请参考[FastQC介绍](#)。使用EIHealth平台创建FastQC应用的详细步骤如下所示：

**步骤1：搭建Docker环境**

**步骤2：制作镜像**

**步骤3：上传镜像**

**步骤4：创建应用**

### 步骤 1：搭建 Docker 环境

1. 搭建Docker环境，您可以任选以下两种方式搭建Docker环境。
  - 使用自己的电脑搭建Docker环境。

- 使用华为云弹性云服务器ECS搭建Docker环境。

本示例中使用华为云弹性服务器ECS，并通过ECS搭建Docker环境。在创建ECS时，可以选择ECS的操作系统。例如，ECS如果为Linux操作系统，可以依次执行如下命令快速安装容器引擎。

```
curl -fsSL get.docker.com -o get-docker.sh
sh get-docker.sh
```

2. 检查安装结果。

执行**docker --version**命令，如果显示如下类似信息，表示Docker安装成功。

图 2-144 Docker 安装成功

```
root@ecs-eihealth:~# docker --version
Docker version 20.10.6, build 370c289
```

## 步骤 2：制作镜像

**方法1：直接下载官方的FastQC镜像。**

执行如下命令下载FastQC镜像。

```
docker pull biocontainers/fastqc:v0.11.5
```

**方法2：通过Dockerfile制作FastQC镜像。**

1. 执行**vi Dockerfile**命令，进入Dockerfile文件中，编写文件。

```
FROM ubuntu:16.04
# FastQC依赖java运行，需安装java环境。安装执行下载、解压缩的软件包
RUN apt-get update && apt-get upgrade -y \
    && apt-get install -y default-jre perl wget zip
# 下载FastQC，解压缩，设置FastQC可执行权限
RUN wget https://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.11.5.zip \
    && unzip fastqc_v0.11.5.zip \
    && rm fastqc_v0.11.5.zip \
    && chmod +x /FastQC/fastqc
# 将FastQC添加到环境变量中
ENV PATH "/FastQC:$PATH"
```

2. 按键盘Esc键，并执行:wq保存退出Dockerfile。
3. 制作镜像。

```
docker build -t fastqc:v0.11.5 .
```

详细的Dockerfile指令请参见[Dockerfile参考](#)。

## 步骤 3：上传镜像

1. [下载命令行工具](#)并进行[初始化配置](#)。
2. 使用**health switch project**命令进入到所需的项目中。

```
# 命令结构
health switch project <project-name>
# 命令示例，例如进入到名为eihealth-user-project的项目中
health switch project eihealth-user-project
```

3. 使用命令行工具上传镜像到EIHealth平台。
  - a. 使用**health docker tag**命令给要上传的镜像打标签。  
health docker tag biocontainers/fastqc:v0.11.5 fastqc:v0.11.5
  - b. 使用**health docker push**命令上传镜像。  
health docker push fastqc:v0.11.5

执行**health docker images**命令查看已有的镜像。

详细的命令介绍请参见“[命令行工具 > 镜像管理命令](#)”章节。

4. 登录EIHealth平台，在镜像列表中查看已上传的镜像。

## 步骤 4：创建应用

1. 登录EIHealth平台，在“项目管理”页面“工具”页签中，单击“新建应用”。

图 2-145 新建应用

| 图标 | 名称                | 标签                           | 源项目                       | 版本     | 创建时间               | 修改时间               | 创建时间               | 操作      |
|----|-------------------|------------------------------|---------------------------|--------|--------------------|--------------------|--------------------|---------|
|    | test-111          | aaaaa.test.test02            | wxy-test                  | 1.0.0  | 2022/02/21 10:0... | 2022/01/27 15:4... |                    | 编辑   删除 |
|    | picard-insertsize | picard.insert.size           | [资产市场]                    | 2.23.3 | 2022/02/14 10:4... | 2022/02/14 10:4... |                    | 编辑   删除 |
|    | bwa-mem           | 高通量二代测序的Clean Reads比对到参考基因组上 | mapping.bwa.NGS           | [资产市场] | 0.7.17             | 2022/02/14 10:0... | 2022/02/14 10:0... | 编辑   删除 |
|    | bamqc             | 评估比对得到的bam文件的质量              | quality.control.bwa       | [资产市场] | 2.0.0              | 2022/02/14 10:0... | 2022/02/14 10:0... | 编辑   删除 |
|    | fastp             | 对测序得到的fastq数据进行质控            | fastp.NGS.quality.control | [资产市场] | 0.20.1             | 2022/02/14 09:5... | 2022/02/14 09:5... | 编辑   删除 |

2. 填写应用的基本信息。

“名称”填写fastqc，“版本”填写v0.11.5.2。“图标”、“标签”、“短描述”、“描述”可选填。

图 2-146 填写基本信息

\* 名称: fastqc

\* 版本: v0.11.5.2

图标:

标签: 请输入标签名称

最近常用标签: 宏观脑科学 | 脑科学 | 功能像fMRI

短描述:

描述:

3. 选择镜像。

单击“选择镜像”，在“自定义镜像”列表中选择fastqc镜像和镜像版本。

4. 依据[FastQC命令说明](#)填写镜像启动命令。

镜像启动命令需要引用输入、输出参数中的变量，并以大括号扩起，以\$符号进行引用。

fastqc软件输入参数填写为input-file、threads，输出参数为output-dir，则镜像启动命令如下所示。

使用-t命令，指定运行所需的线程数量。-o命令，指定存放输出结果的文件夹。输入文件夹已在填写参数时指定。

```
fastqc -t ${threads} -o ${output-dir} ${input-file}
```

5. 选择“X86”CPU架构，CPU需求建议0.2起。GPU类型选择“无”。

6. 按需填写内存大小，单位为GB。FastQC运行中所需内存大小依赖于输入数据大小，建议至少1GB。计算节点标签可将作业运行于指定的资源节点上，标签设置方法请参见[计算资源标签管理](#)。

图 2-147 CPU、内存、GPU

CPU架构: X86 | ARM

Memory需求: 1 GB

GPU类型: 无 | GPU | D310 | D910

计算节点标签:

CPU需求: 0.2

GPU需求: 0

7. 填写参数。

- a. 通过阅读[FastQC命令说明](#)，了解命令。

图 2-148 FastQC 命令

```
my $result = GetOptions('version' => \$version,
                        'help' => \$help,
                        'quiet' => \$quiet,
                        'nogroup' => \$nogroup,
                        'expgroup' => \$expgroup,
                        'outdir' => \$outdir,
                        'extract!' => \$unzip,
                        'formats=' => \$format,
                        'threads=i' => \$threads,
                        'kmer=i' => \$kmer_size,
                        'casava' => \$casava,
                        'nano' => \$nano,
                        'nofilter' => \$nofilter,
                        'contaminants=' => \$contaminant,
                        'adapters=' => \$adapter,
                        'limits=' => \$limits,
                        'dir=s' => \$temp_directory,
                        'java=' => \$java_bin,
                        'min_length=i' => \$min_length,
                        );

The options for the program as follows:
-h --help      Print this help file and exit
-v --version   Print the version of the program and exit
-o --outdir    Create all output files in the specified output directory.
               Please note that this directory must exist as the program
               will not create it. If this option is not set then the
               output file for each sequence file is created in the same
               directory as the sequence file which was processed.
--casava       Files come from raw casava output. Files in the same sample
               group (differing only by the group number) will be analysed
               as a set rather than individually. Sequences with the filter
               flag set in the header will be excluded from the analysis.
               Files must have the same names given to them by casava
               (including being gzipped and ending with .gz) otherwise they
               won't be grouped together correctly.
--nano         Files come from nanopore sequences and are in fast5 format. In
               this mode you can pass in directories to process and the program
               will take in all fast5 files within those directories and produce
               a single output file from the sequences found in all files.
--nofilter     If running with --casava then don't remove read flagged by
               casava as poor quality when performing the QC analysis.
```

- b. 填写所需的输入参数。

图 2-149 输入参数

| 参数名称       | 数据类型   | 参数模式 | 描述 | 必传                       | 开发                       | 其他                                  | 操作                                |
|------------|--------|------|----|--------------------------|--------------------------|-------------------------------------|-----------------------------------|
| input-file | File   |      |    | <input type="checkbox"/> | <input type="checkbox"/> | 默认值 <input type="text"/>            | <input type="button" value="清除"/> |
| threads    | String |      |    | <input type="checkbox"/> | <input type="checkbox"/> | 默认值 <input type="text" value="30"/> | <input type="button" value="清除"/> |

- c. 填写所需的输出参数。

因镜像启动命令中指定了输出参数，设置输出参数时，需勾选“必传”，并填写“默认值”。例如，输出结果默认存放在fastqc\_output文件夹中。

图 2-150 输出参数

| 参数名称       | 数据类型      | 参数模式 | 描述 | 必传                                  | 其他                                              | 操作                                |
|------------|-----------|------|----|-------------------------------------|-------------------------------------------------|-----------------------------------|
| output-dir | Directory |      |    | <input checked="" type="checkbox"/> | 默认值 <input type="text" value="/fastqc_output"/> | <input type="button" value="清除"/> |

8. 单击“立即创建”，完成fastqc应用的创建。

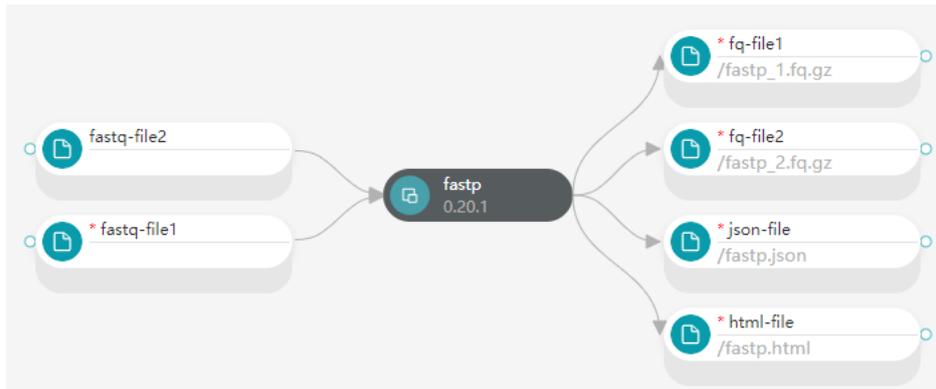
创建完成后的应用，将显示在应用列表中，您可以使用该应用创建分析作业。

## 2.11.8 新建流程

### 流程说明

分析流程至少由一个应用组成，在多个应用构成的流程中，一个应用的输出作为另一个应用的输入，流程中的各个应用由其前后顺序形成完整的计算 workflow。

- EIHealth中的流程由应用搭建形成，应用包含了数据的输入、输出等参数定义。



- 应用呈现的信息在[创建应用](#)过程中定义，包含参数名称、数据类型、描述、默认值等。



- 单击参数名称，可查看参数的详细信息。



- 单击输入参数  图标，可指定输入数据路径。输出数据路径在[新建流程](#)和[运行分析作业](#)时可指定。

## 新建流程

- 单击“新建流程”，进入流程设计器页面。  
流程设计器是一种用于创建、查看、修改流程的图形化工具，详细介绍请参见[流程设计器](#)。新建流程时，自动保存默认打开。

图 2-151 新建流程

| 名称                           | 标签  | 源项目      | 版本    | 创建者           | 修改时间                | 创建时间                | 操作             |
|------------------------------|-----|----------|-------|---------------|---------------------|---------------------|----------------|
| AAA_123_b6                   | ... | wyq-test | 1.2   | et_ehealth... | 2022/03/23 18:13:43 | 2022/03/23 18:13:43 | 启动作业   编辑   删除 |
| _123-mub112ATEE-1<br>summary | ... | wyq-test | 1.0.0 | et_ehealth... | 2022/03/18 17:05:37 | 2022/02/24 09:51:08 | 启动作业   编辑   删除 |

2. 填写流程参数信息，填写完成后单击“确定”。
  - a. 填写流程基本参数，包括“流程名称”、“版本”、“标签”、“短描述”和“描述”。  
“流程名称”和“版本”为必填项，其他参数可选填。

图 2-152 流程参数

**流程设置** ✕

\* 流程名称

\* 版本

标签

最近常用标签: 宏观脑科学 aaaa 猴脑

短描述

描述

超时时间  mins

输出路径

- b. 填写“超时时间”、“输出路径”。  
为可选参数，基于流程创建分析作业时，该参数可重新定义。
      - “超时时间”指作业运行时间超过设置时间时，认为超时，默认1440分钟，最大可设置为144000分钟，即作业运行至多100天。
      - “输出路径”指运行分析作业时的输出结果存放路径。  
例如，gene-assets项目中的output文件夹，输出路径格式为/gene-assets/output。

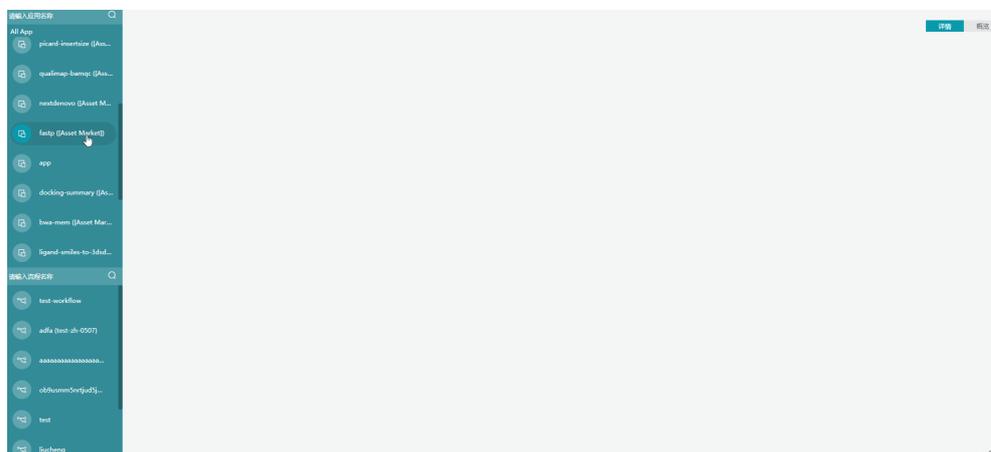
图 2-153 流程参数

超时时间  mins

输出路径

3. 将左侧应用列表中的应用，拖拽至画布中。对于由多个应用构成的流程，通过连接线指定输入、输出关系进行链接。  
鼠标移动至应用中，会出现一个“空心圆”，拖动“空心圆”，将连接线移动到目标应用上。

图 2-154 搭建流程



### 说明

只有当应用的输入、输出参数类型相同时，才可以进行连接。例如，将app1和app2连接起来，app1的输出，app2的输入均为file类型。

4. 单击应用图标，弹出编辑、删除、复制图标，可对应用执行相应操作。

图 2-155 编辑应用



5. 单击  按钮，进入参数设置页面，可查看修改参数。
  - 名称：应用的原始名称。
  - 版本：应用版本。
  - 显示名称：应用在流程中的显示名称，可修改。
  - 镜像启动命令：镜像启动的命令。可在右上角放大查看，在右下角下拉查看。
  - 源项目：显示应用隶属的项目。
  - 输入参数：由用户在创建应用时定义。
  - 输出路径：输出数据的存放路径，可修改。
  - 高级参数：CPU需求、Memory需求、GPU类型、GPU需求，请按照使用需求进行设置；CPU架构、计算节点标签，不可修改。GPU包含NAIDIA架构GPU

和自研的D310+ARM。选择GPU资源时，需要您在购买平台时包含了GPU资源，并且需要应用支持GPU运行。选择D310+ARM时，需要应用支持ARM环境运行，并在创建应用时，镜像系统为ARM类系统。

### 📖 说明

运行分析作业时，流程中的每一个应用称之为一个任务（Task），在编排流程时，如果“输出路径”修改为空，实际作业输出结果依然默认存在该路径。最终输出结果路径按照“流程输出路径/作业名称-UUID/Task输出路径/应用输出路径”规则生成。

6. 单击界面上方按钮，保存流程。并提示“流程保存成功”。  
创建好的流程，将显示在流程列表中。

## 2.11.9 流程设计器

分析流程至少由一个应用构成，流程中的各个应用由其前后顺序形成完整的计算 workflow，一个应用的输出作为另一个应用的输入。

流程设计器是一种用于创建、查看、修改流程的图形化工具。借助流程设计器，您可以拖拽工具到画布中，可视化链接各应用，指定应用的先后顺序。

### 流程设计器界面

流程设计器界面由工具栏、资源栏和画布三部分构成。

图 2-156 流程设计器界面

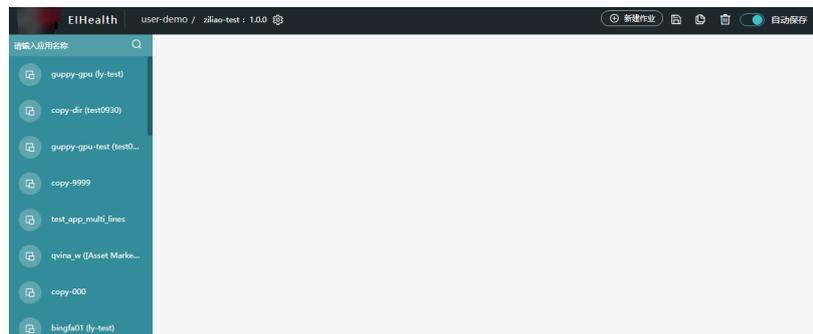


表 2-21 界面说明

| 区域  | 说明                                                |
|-----|---------------------------------------------------|
| 工具栏 | 上方的工具栏显示设计器的快捷控制操作。<br>由设置、新建作业、保存、另存为、删除、自动保存构成。 |
| 资源栏 | 左侧的资源栏显示项目中可以使用的应用和流程。                            |

| 区域 | 说明                                                                                                                                                                                                                                                                                                                                                                                                 |
|----|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 画布 | <p>中间区域为搭建流程的操作界面。</p> <ul style="list-style-type: none"> <li>可以将应用拖拽至画布中，并进行编排创建。</li> <li>可以将流程放置到画布中进行编排修改。</li> <li>可通过“概览”展示流程的结构图。</li> <li>可通过界面右下角快速定位  按钮使流程显示在画布中央。</li> <li>可通过界面右下角规整排列  按钮使流程连线排布整齐。</li> </ul> |

## 2.11.10 导入流程

导入流程是将隶属于其他项目中流程导入至本项目中，流程所依托的应用和镜像会同步导入。

使用“导入流程”功能，用户需是其他项目中的成员，且为其他项目的“所有者”或“管理员”。

1. 单击“导入流程”，进入导入流程页面。

图 2-157 导入流程



2. 选择需要引用的项目以及项目中的流程，选择流程的版本。“导入流程名称”是选填项，可以使用原有名称，或自定义。

图 2-158 导入流程



3. 单击“确定”，导入流程。

## 2.11.11 上传流程

通过在本地修改流程的yaml模板，上传流程至项目中。

1. 获取流程yaml模板。
  - 单击“上传流程”，在弹出的页面中下载yaml示例文件。

- 使用命令行工具，执行**health get workflow -s**命令获取创建流程的yaml模板，复制模板并保存到本地。可以保存成.yaml或.txt文件，保存为.txt文件时，其内容需为yaml格式。

## 2. 修改流程yaml模板。

在本地编写模板，模板中的参数与在EIHealth平台**新建流程**过程一致。

修改yaml模板时，详细的命令和参数请参见[流程配置文件说明](#)。

# 详情说明可参考API文档中流程管理- 创建流程

```

workflow:
  id: # 流程id
  name: 'demo-workflow' # 流程名称，取值范围[1,56]，以大小写字母开头，允许出现中划线(-)、下划线(_)、大小写字母和数字，且必须以大小写字母或数字结尾。更新流程时，流程名称不支持修改。
  version: '1.0.0' # 流程版本，取值范围[1,24]，以小写字母或数字或大写字母开头，允许出现中划线，必须以大小写字母或数字结尾。更新流程时，流程版本不支持修改。
  summary: 'summary' # 流程简述 取值范围[0,128]
  description: 'description' # 流程描述 取值范围[0,65535]，后续支持markdown文本
  labels: # 流程标签，取值范围[0,5]，标签名称长度为1~32，支持中文、字母、数字、空格、下划线和中划线，且不能以空格开头或者结尾。
    - labelA
    - labelB
  timeout: 1440 # 流程超时时间，取值范围[1,144000]，单位分钟，默认1440
  output_dir: # 流程的当前工作目录，默认为根目录，用户可显示指定，必须以/开头，结尾不能有/。
  tasks: # 流程中子任务的描述信息，子任务数量取值范围:[1,128]
    - task_name: app1-1 # 子任务实际名称，取值范围[1,32]，只能以大小写字母开头，由大小写字母、数字、中划线(-)、下划线(_)组成，以大小写字母或数字结尾。
      app_id: 'demo-app::1.0.0:gwj-test-01' # 应用id，取值范围[1,135]，正则先不能有中文，两种格式。特殊id，采用{app_name}::{app_version}::{src_project_name}格式，用于手动创建场景；
      # 其他场景，app_id为系统分配的唯一标识
      display_name: '步骤1' # 流程的子任务展示名称，最大长度64
      resources: # 子任务的资源配额
        cpu: '0.1C' # cpu申请使用量，取值范围[0.1-128]，单位C，支持一位小数。对于应用，不填默认1C；对于流程和作业，不填默认使用前一级的配置，填值会覆盖更新。覆盖关系：作业->流程->应用
        memory: '0.1G' # 内存申请使用量，取值范围[0.1-3072]，单位G，支持一位小数。对于应用，不填默认1G；对于流程和作业，不填默认使用前一级的配置，填值会覆盖更新。覆盖关系：作业->流程->应用
        gpu_type: "" # gpu架构类型，取值范围 '|GPU|Snt9|D310'。对于流程和作业，不填默认使用前一级的配置，填值会覆盖更新。覆盖关系：作业->流程->应用
        gpu: '0' # gpu申请使用量，取值范围[0-16]，仅支持整数，Snt9有特殊约束，申请数量需要是0,1,2,4,8。对于应用，不填默认0；对于流程和作业，不填默认使用前一级的配置，填值会覆盖更新。
      # 覆盖关系：作业->流程->应用
      output_dir: # 子任务的输出存放路径，用户可显示指定，必须以/开头，结尾不能有/。
      inputs:
        - name: 'input-dir' # 子任务的参数名称，长度为[1,32]，以小写字母开头，允许出现中划线(-)、小写字母和数字，且必须以小写字母或数字结尾。需要和已有应用的参数名称一致。
          values: # 子任务的参数数值，根据参数类型进行合法性校验
            - 'gwj-test-01:/test'
        - task_name: app1-2
          app_id: 'demo-app::1.0.0'
          display_name: '步骤2'
          resources:
            cpu: '0.1C'
            memory: '0.1G'
            gpu_type: ""
            gpu: '0'
            output_dir:
            inputs:
              - name: 'input-dir'
                values: # 子任务的参数数值，格式为${task_name.output_name}，表示连线关系，即app1-2需要app1-1成功执行后才会运行
                  - '${app1-1.output-dir}'
  
```

## 3. 上传流程模板。

单击“上传流程”，在弹出的页面中上传yaml模板。

## 2.11.12 发布流程

您可以将创建的流程发布到资产市场，供其他项目订阅和使用。

发布资产前，需要先设置商标。详细可参考[商标设置](#)。

1. 在“工具”页面选择“流程”。
2. 选择需要发布的流程，单击“操作”列“更多”>“发布”。
3. 在发布页面，填写发布流程的相关信息。

表 2-22 发布流程参数说明

| 参数   | 说明                                                                                                                                  |
|------|-------------------------------------------------------------------------------------------------------------------------------------|
| 名称   | 发布流程的名称。名称长度为1-56，只允许出现中划线、下划线、字母和数字。                                                                                               |
| 展示名称 | 发布成功后在组织共享中展示的名称。若不填写，展示的名称和流程名称相同。发布成功的流程不支持修改展示名称。                                                                                |
| 版本   | 发布流程的版本。<br>版本长度必须为1-24个字符，以字母或数字开头（多个字符可以连续），中间只能包含-特殊符号（多个特殊符号不能连续），以字母或数字结尾。                                                     |
| 类型   | 流程。                                                                                                                                 |
| 标签   | 设置发布流程的标签。<br>最多支持设置5个标签。                                                                                                           |
| 短描述  | 为发布的流程设置的简要描述。                                                                                                                      |
| 图片   | 展示在资产市场的图片。要求如下： <ul style="list-style-type: none"> <li>● 格式：svg,png,jpg,jpeg</li> <li>● 尺寸：16 * 9</li> <li>● 大小：不超过20KB</li> </ul> |
| 描述   | 针对此流程的相关描述。                                                                                                                         |

4. 单击“立即发布”。  
系统跳转到“我的发布”页面，可以查看发布的流程状态。当流程发布成功后，可以在“资产市场>组织共享”中查看发布的流程。

## 说明

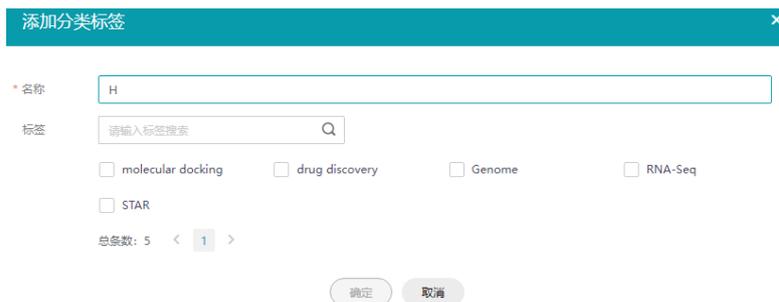
- 已发布的流程不能直接编辑和删除，需下线后才能编辑和删除。
- 发布中的流程可以取消，取消后状态为发布失败。
- 已发布的流程可以订阅和下线。
- 同一个发布者可以发布同名称，同类型的流程。若版本不同，则合并。用户在使用时，可以选择对应的版本号进行订阅或者下线。若版本相同，则更改版本后发布。
- 已发布的应用，可以在“资产市场>我的发布”中进行下线。下线后已经订阅的应用不可以使用。

## 2.11.13 添加分类标签

为了方便分类查看创建的应用和流程，您可以使用分类标签功能，筛选呈现带有相同标签的应用或流程。

1. 在“工具”页面左侧，单击 **+** 图标新建分类标签。
2. 在添加分类标签弹窗中，输入“名称”，勾选需要呈现的标签名称。最多可以勾选5个标签。

图 2-159 添加分类标签



3. 勾选完成后单击“确定”。
4. 页面自动筛选出带有所选标签的应用或流程。

图 2-160 添加完成



## 2.11.14 下载应用或流程

1. 在“工具”页面选择“流程”或“应用”。
2. 选择需要发布的流程，单击“操作”列“更多”>“下载”。

图 2-161 下载应用



下载内容不包含应用和流程的id，所有引用关系通过名称和版本来指定。

## 2.12 作业管理

### 2.12.1 分析作业管理简介

分析作业依托于流程运行，需要您先[创建流程](#)，再基于流程创建分析作业。

在“项目管理”页面“作业”页签中，以列表形式展示了项目中运行的分析作业和运行状态。您可以查看分析作业名称、标签、所使用的流程名称、版本、运行状态、创建者、创建时间、完成时间和总耗时。并可对作业执行重试、取消、删除、克隆、导出操作。对于列表中的作业，支持通过作业名称、流程名称、标签、创建者、状态、创建时间和完成时间进行快速搜索。

如果需要批量创建分析作业，您可以创建自动作业。

#### 说明

- 当作业执行超时，作业状态显示失败，task状态为运行中。后续task运行成功或失败，作业状态都显示失败。可参考[作业投递后处于运行中，运行过程正常，但是最后超时失败](#)进行处理。
- 如果性能加速节点不可用或作业配额不足，重试作业时，超过配额的作业会重试失败。

图 2-162 分析作业

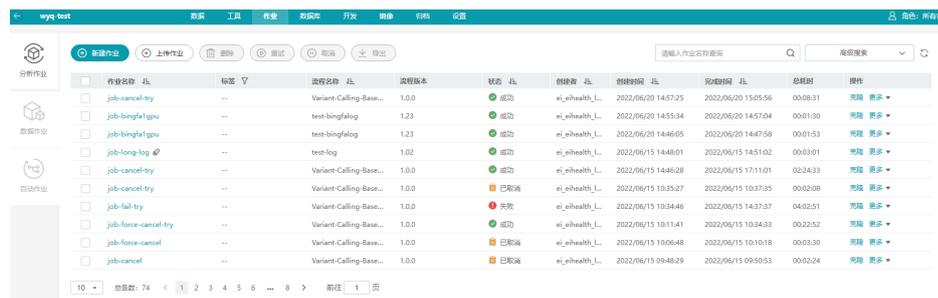


图 2-163 作业搜索



### 2.12.2 创建分析作业

创建分析作业有以下几种不同的方式，您可以任选其中一种方式。

- 通过“新建流程”时创建
- 通过“作业”页面创建
  - •新建作业
  - •克隆作业
- 通过“工具”页面创建
- 通过“上传作业”创建

#### 📖 说明

如果在创建应用时打开了“并发”开关，可以设置多个参数值，批量执行作业。

可通过在“系统资源 > 计算资源”页面将计算节点的标签设置为copy-in，实现作业运行于指定的计算节点中。对于“不可调度”状态的计算节点，将无法在此节点运行作业。

## 通过“新建流程”时创建

1. 在“工具”页面，单击“新建流程”，填写流程信息。请参考[新建流程](#)操作。
2. 流程搭建完成后，单击“新建作业”。

图 2-164 新建作业



3. 在新建作业弹窗中单击“确定”。

图 2-165 新建作业



4. 在弹出的作业设置页面，填写作业信息。“作业名称”、“标签”、和“描述”，选择需要执行的流程，设置“输出路径”、“优先级”、“计算节点标签”、“超时时间”、“加速类型”。设置完成后单击“确定”。
  - 基本信息：包含作业名称、标签、描述。
  - 流程名称：显示已创建的流程名称，不可编辑。
  - 输出路径：存放输出结果的路径，格式以/开头。例如项目中的output文件夹，输出路径可设置为/output。不填写路径时，默认以“作业名-UUID”格式生成输出路径。
  - 优先级：运行优先级，分为0~9级，优先级高的作业会被优先执行。默认值为0。

#### 📖 说明

如果当前投递的IO加速的作业，超过sfs总的作业配额数，那么超出部分的作业会按照投递时间顺序，从最新投递的向前执行，不会按照优先级进行运行。

- 计算节点标签：作业会调度到含有相应标签的计算节点上。

当应用也配置了标签，如果应用和作业的计算节点标签在同一计算节点上，则应用调度至该计算节点上；应用和它的作业，不管节点标签是否一致，都会被调度到应用的节点标签所对应的计算节点上。

如果设置了不存在的计算节点标签，作业会进入等待，直至配置了相应的标签。

- 超时时间：作业运行时间超过设置时间时，认为超时，默认1440分钟，最大可设置为144000分钟，即作业运行至多100天。

- 加速类型：

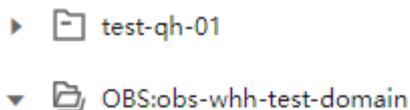
加速效率：IO加速>本地盘加速>无

- 无：作业运行于OBS中，不使用加速。
- IO加速：IO加速使用弹性文件服务（SFS）提供高性能的数据读写，作业运行时，会将非最终结果的数据存储在SFS中用以提高任务运行效率，作业执行完成后会清理释放SFS空间。对于涉及频繁读写场景的任务建议开启IO加速，开启前需要先[购买性能加速](#)。
- 本地盘加速：使用计算节点的本地盘进行加速。使用本地盘加速时，需保证购买的[计算节点](#)带有“数据盘”。OBS桶中的数据不支持本地盘加速，使用OBS桶中数据用于本地盘加速，可能会导致作业运行失败。

#### 📖 说明

- 使用OBS桶中的数据投递作业时，当数据大于40G时作业会投递失败。
- 使用OBS桶中的数据投递作业时，作业将数据copy至云硬盘后，数据变大了4096byte，导致作业投递失败。

图 2-166 “OBS” 标签代表数据引用来源为 OBS 桶





6. 单击界面上方“启动作业”按钮，可以直接基于该流程创建分析作业。

## 通过“作业”页面创建

- 新建作业

- 单击“新建作业”，在弹出的新建作业页面，填写作业信息。“作业名称”、“标签”、和“描述”，选择需要执行的流程，设置“输出路径”、“优先级”、“计算节点标签”、“超时时间”、“加速类型”。
  - 基本信息：包含作业名称、标签、描述。
  - 流程名称：选择需要执行的流程。
  - 输出路径：存放输出结果的路径，格式以/开头。例如项目中的output文件夹，输出路径可设置为/output。不填写路径时，默认以“作业名-UUID”格式生成输出路径。
  - 优先级：运行优先级，分为0~9级，优先级高的作业会被优先执行。默认值为0。

### 说明

如果当前投递的IO加速的作业，超过sfs总的作业配额数，那么超出部分的作业会按照投递时间顺序，从最新投递的向前执行，不会按照优先级进行运行。

- 计算节点标签：作业会调度到含有相应标签的计算节点上。  
当应用也配置了标签，如果应用和作业的计算节点标签在同一计算节点上，则应用调度至该计算节点上；应用和它的作业，不管节点标签是否一致，都会被调度到应用的节点标签所对应的计算节点上。  
如果设置了不存在的计算节点标签，作业会进入等待，直至配置了相应的标签。
- 超时时间：作业运行时间超过设置时间时，认为超时，默认1440分钟，最大可设置为144000分钟，即作业运行至多100天。
- 加速类型：
  - 加速效率：IO加速>本地盘加速>无
  - 无：作业运行于OBS中，不使用加速。
  - IO加速：IO加速使用弹性文件服务（SFS）提供高性能的数据读写，作业运行时，会将非最终结果的数据存储在SFS中用以提高任务运行效率，作业执行完成后会清理释放SFS空间。对于涉及频繁读写场景的任务建议开启IO加速，开启前需要先[购买性能加速](#)。
  - 本地盘加速：使用计算节点的本地盘进行加速。使用本地盘加速时，需保证购买的**计算节点**带有“数据盘”。OBS桶中的数据不支持本地盘加速，使用OBS桶中数据用于本地盘加速，可能会导致作业运行失败。

图 2-169 “OBS” 标签代表数据引用来源为 OBS 桶

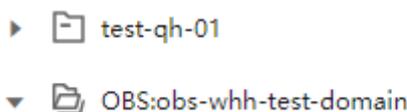


图 2-170 作业设置

- b. 填写信息确认无误后，单击“确定”，进入流程设计器页面。
  - c. 单击输入参数  图标，设置输入数据。
  - d. 单击应用图标，弹出编辑图标，单击  按钮，设置应用参数。
  - e. 参数确认无误后，单击界面上方“启动作业”按钮，可以直接基于该流程创建分析作业。
- 克隆作业  
克隆已有的分析作业，作业的参数及流程一并克隆到流程设计器中。克隆作业后，请修改作业的输出数据路径，否则会有覆盖原有作业数据风险。
- a. 执行克隆操作有两种方式，请选择任一方式执行。

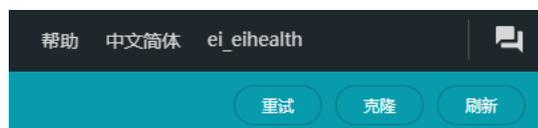
- 在分析作业列表中，单击“操作”列“克隆”，进入流程设计器页面。

图 2-171 克隆作业



- 在分析作业列表中，单击作业名称，进入详情页。单击右上角“克隆”，进入流程设计器页面。

图 2-172 克隆作业



- b. 单击输入参数  图标，设置输入数据。
- c. 单击应用图标，弹出编辑图标，单击  按钮，设置应用参数。
- d. 参数确认无误后，单击界面上方“启动作业”按钮，可以直接基于该流程创建分析作业。

## 通过“工具”页面创建

1. 在“工具”页面，流程列表中，单击“操作”列“启动作业”。在弹出的作业设置页面，填写作业信息。“作业名称”、“标签”、和“描述”，设置“输出路径”、“优先级”、“计算节点标签”、“超时时间”、“加速类型”。
  - 基本信息：包含作业名称、标签、描述。
  - 输出路径：存放输出结果的路径，格式以/开头。例如项目中的output文件夹，输出路径可设置为/output。  
不填写路径时，默认以“作业名-UUID”格式生成输出路径。
  - 优先级：运行优先级，分为0~9级，优先级高的作业会被优先执行。默认值为0。

### 说明

如果当前投递的IO加速的作业，超过sfs总的作业配额数，那么超出部分的作业会按照投递时间顺序，从最新投递的向前执行，不会按照优先级进行运行。

- 计算节点标签：作业会调度到含有相应标签的计算节点上。  
当应用也配置了标签，如果应用和作业的计算节点标签在同一计算节点上，则应用调度至该计算节点上；应用和它的作业，不管节点标签是否一致，都会被调度到应用的节点标签所对应的计算节点上。  
如果设置了不存在的计算节点标签，作业会进入等待，直至配置了相应的标签。
- 超时时间：作业运行时间超过设置时间时，认为超时，默认1440分钟，最大可设置为144000分钟，即作业运行至多100天。
- 加速类型：
  - 加速效率：IO加速>本地盘加速>无
  - 无：作业运行于OBS中，不使用加速。
  - IO加速：IO加速使用弹性文件服务（SFS）提供高性能的数据读写，作业运行时，会将非最终结果的数据存储在SFS中用以提高任务运行效率，作业执行完成后会清理释放SFS空间。对于涉及频繁读写场景的任务建议开启IO加速，开启前需要先[购买性能加速](#)。
  - 本地盘加速：使用计算节点的本地盘进行加速。使用本地盘加速时，需保证购买的[计算节点](#)带有“数据盘”。OBS桶中的数据不支持本地盘加速，使用OBS桶中数据用于本地盘加速，可能会导致作业运行失败。

图 2-173 “OBS” 标签代表数据引用来源为 OBS 桶

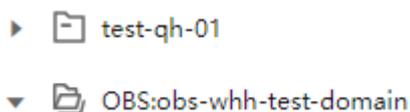


图 2-174 启动作业



2. 作业信息填写完成后，单击“确定”，进入流程设计器页面。
3. 单击输入参数  图标，设置输入数据。
4. 单击应用图标，弹出编辑图标，单击  按钮，设置应用参数。
5. 参数确认无误后，单击界面上方“启动作业”按钮，可以直接基于该流程创建分析作业。

## 通过“上传作业”创建

通过在本地修改作业的yaml模板，运行分析作业。

1. 获取作业yaml模板。
  - 在“项目管理 > 作业”页签中，单击“上传作业”，在弹出的页面中下载yaml示例文件。
  - 使用命令行工具，执行 `health get job -s` 命令获取创建作业的yaml模板，复制模板并保存到本地。可以保存成.yaml或.txt文件，保存为.txt文件时，其内容需为yaml格式。

2. 修改作业yaml模板。

在本地编写模板，模板中的参数与在EIH Health平台创建作业过程一致。

修改yaml模板时，详细的命令和参数请参见[作业配置文件说明](#)。

```
# 详情说明可参考API文档中作业管理-创建作业
job:
  name: demo-job # 作业名称, 取值范围: [1,63], 以小写字母开头, 允许出现中划线(-)、数字和小写字母, 且必须以小写字母或数字结尾
  description: description # 作业描述,取值范围: 输入字符最大长度为255
  priority: 0 # 作业的优先级,取值范围[0,9], 0最低, 默认数值0
  timeout: 1440 # 作业执行超时时长, 取值范围: [1, 144000], 单位: 分钟, 默认数值1440
  output_dir: # 作业存储目录, 不指定则在workflow的工作目录下生产job同名子目录, 指定则已指定路径为准, 必须以/开头, 结尾不能有/
  workflow_id: demo-workflow::1.0.0::gwj-test-01 # 作业依赖的组件id, 组件当前仅支持流程, 取值范围[1,135], 支持大小写字母和数字。目前支持两种格式, 特殊id: {流程名称}::{流程版本}::{源项目名称}; 正常id: 流程id
  io_acc_id: # IO加速实例id, 为空则不开启IO加速, 设置相应加速包id将会开启加速, 当id设置为: auto_schedule, 则会自动调度加速包, 当id设置为 local_disk, 则开启本地盘加速。
  node_labels: # 计算节点标签 标签个数取值范围[0-1],单个节点标签最大长度56个字符, 以字母或数字开头, 可以包含下划线(_)、中划线(-)、点(.)、字母和数字, 且必须以字母或数字结尾
  - health.label
  tasks:
  - task_name: app1-1 # 子任务实际名称, 取值范围[1,32], 只能以大小写字母开头, 由字母、数字、中划线(-)、下划线(_)组成, 以大小写字母或数字结尾。需要和已有子任务的名称一致。
    inputs:
    - name: input-dir # 子任务的输入参数信息
      values: # 子任务的参数数值, 根据参数类型进行合法性校验
      - 'gwj-test-01:/test'
    resources:
    cpu: 0.1C # 子任务的资源配额
      # cpu申请使用量, 取值范围[0.1-128], 单位C, 支持一位小数。
      对于应用, 不填默认1C; 对于流程和作业, 不填默认使用前一级的配置, 填值会覆盖更新。覆盖关系: 作
```

```

业->流程->应用
  memory: 0.1G # 内存申请使用量，取值范围[0.1-3072]，单位G，支持一位
小数。对于应用，不填默认1G；对于 流程和作业，不填默认使用前一级的配置，填值会覆盖更新。覆盖关
系：作业->流程->应用
  gpu_type: " # gpu架构类型，取值范围 '|GPU|Snt9|D310。对于流程和作
业，不填默认使用前一级的配置，填值会覆盖更新。覆盖关系：作业->流程->应用
  gpu: '0' # gpu申请使用量，取值范围[0-16]，仅支持整数，Snt9有特殊约
束，申请数量需要是0,1,2,4,8 。对于应用，不填默认0；对于流程和作业，不填默认使用前一级的配置，填
值会覆盖更新。
 # 覆盖关系：作业->流程->应用
  io_acc_type #子任务使用的IO加速类型，取值范围SFS、EVS。为空表示不使
用加速。SFS表示使用io加速，EVS表示使用本地盘加速
- task_name: app1-2
inputs:
- name: input-dir
  values:
  - '${app1-1.output-dir}'
resources:
  cpu: 0.1C
  memory: 0.1G
  gpu_type: "
  gpu: '0'
  io_acc_type: SFS
    
```

### 3. 上传并运行作业。

单击“上传作业”，在弹出的页面中上传yaml模板，上传成功后，即可运行作业。

## 2.12.3 创建自动作业

针对存在需要批量创建分析作业的场景，您可以选择创建自动作业。

在“自动作业”页签，您可以查看已创建的自动作业，包含作业名称、状态、数据表、创建者、创建时间；在操作列，您可以对已创建的自动作业执行启动、编辑、删除操作，运行中的作业可以执行停止操作。

图 2-175 自动作业

| 作业名称     | 状态 | 数据表             | 创建者          | 创建时间                | 操作           |
|----------|----|-----------------|--------------|---------------------|--------------|
| test-bug | 停止 | test-bug        | ei_ehealth_L | 2022/06/20 11:38:12 | 启动   编辑   删除 |
| test123  | 失败 | long-double-150 | ei_ehealth_L | 2022/06/15 16:44:34 | 启动   编辑   删除 |
| sdd      | 失败 | test-quanta     | ei_ehealth_L | 2022/06/14 14:27:49 | 启动   编辑   删除 |
| 222      | 停止 | a0613           | ei_ehealth_L | 2022/06/13 17:07:06 | 启动   编辑   删除 |
| test-ed8 | 停止 | string-out      | ei_ehealth_L | 2022/06/13 14:42:25 | 启动   编辑   删除 |
| aada-fid | 失败 | alltype         | ei_ehealth_L | 2022/06/13 10:49:58 | 启动   编辑   删除 |
| test-111 | 停止 | vstring         | ei_ehealth_L | 2022/06/10 16:12:26 | 启动   编辑   删除 |
| heKa-0   | 停止 | ced             | ei_ehealth_L | 2022/06/10 16:00:49 | 启动   编辑   删除 |
| iter6    | 停止 | iter6           | ei_ehealth_L | 2022/06/10 09:35:59 | 启动   编辑   删除 |
| test123  | 停止 | multi           | ei_ehealth_L | 2022/06/09 18:55:36 | 启动   编辑   删除 |

## 前提条件

- 在“工具”中完成[流程创建](#)。
- 在“数据库”中完成样本集的[数据库创建](#)。数据库的每一行对应一个任务，如果满足触发条件，就会在平台自动运行，并在分析任务列表添加一条记录。

## 创建自动作业

- 在“作业”页面左侧，单击“自动作业”页签。
- 在页面左上角单击“新建自动作业”。

3. 设置作业基本信息，包括“名称”、“描述”、“数据表”、“流程名称”、“作业状态更新”。
  - “作业状态更新”中可选“选择已有列”或“新建一列”。
    - 选择“选择已有列”时，需要选择已有列名，已有列名必须为string类型，可以搜索到，并且非主键。  
并设置是否“清空作业状态更新列”，状态更新列的值为空的数据才会投递作业。
    - 选择“新建一列”，需要输入新建的列名。

图 2-176 设置作业基本信息

4. 定义触发器。使用触发器来筛选需要运行作业的数据，并且满足状态更新列的值为空的数据才会被投递作业。  
单击“添加条件”，在数据列名、判断模式、值中定义投递作业的触发条件。数据列名不能选择作业状态更新列，并且列名可搜索。

图 2-177 定义触发器

5. 填写作业参数配置。填写完成后，单击“下一步：设置流程参数”。
  - 作业名称：可选择“数据库表”、“自定义”或“自动生成”。
  - 标签：设置作业标签。
  - 描述：根据需要填写。
  - 输出路径：可选择“数据库表”和“自定义”。选择“自定义”时，存放输出结果的路径，格式以/开头。例如项目中的output文件夹，输出路径可设置为/output。不填写路径时，默认以“作业名-UUID”格式生成输出路径。
  - 优先级：运行优先级，分为0~9级，优先级高的作业会被优先执行。默认值为0。
  - 计算节点标签：作业会调度到含有相应标签的计算节点上。  
当应用和作业都配置了标签，如果应用和作业的计算节点标签在同一计算节点上，则应用调度至该计算节点上；应用和它的作业，不管节点标签是否一致，都会被调度到应用的节点标签所对应的计算节点上。  
如果设置了不存在的计算节点标签，作业会进入等待，直至配置了相应的标签。
  - 超时时间：作业运行时间超过设置时间时，认为超时，默认1440分钟，最大可设置为144000分钟，即作业运行至多100天。
  - 加速类型：  
加速效率：IO加速>本地盘加速>无

- 无：作业运行于OBS中，不使用加速。
- IO加速：IO加速使用弹性文件服务（SFS）提供高性能的数据读写，作业运行时，会将非最终结果的数据存储在SFS中用以提高任务运行效率，作业执行完成后会清理释放SFS空间。对于涉及频繁读写场景的任务建议开启IO加速，开启前需要先[购买性能加速](#)。
- 本地盘加速：使用计算节点的本地盘进行加速。使用本地盘加速时，需保证购买的[计算节点](#)带有“数据盘”。OBS桶中的数据不支持本地盘加速，使用OBS桶中数据用于本地盘加速，可能会导致作业运行失败。

#### 📖 说明

- 使用OBS桶中的数据投递作业时，当数据大于40G时作业会投递失败。
- 使用OBS桶中的数据投递作业时，作业将数据copy至云硬盘后，数据变大了4096byte，导致作业投递失败。

图 2-178 “OBS” 标签代表数据引用来源为 OBS 桶

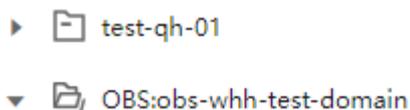


图 2-179 作业参数配置

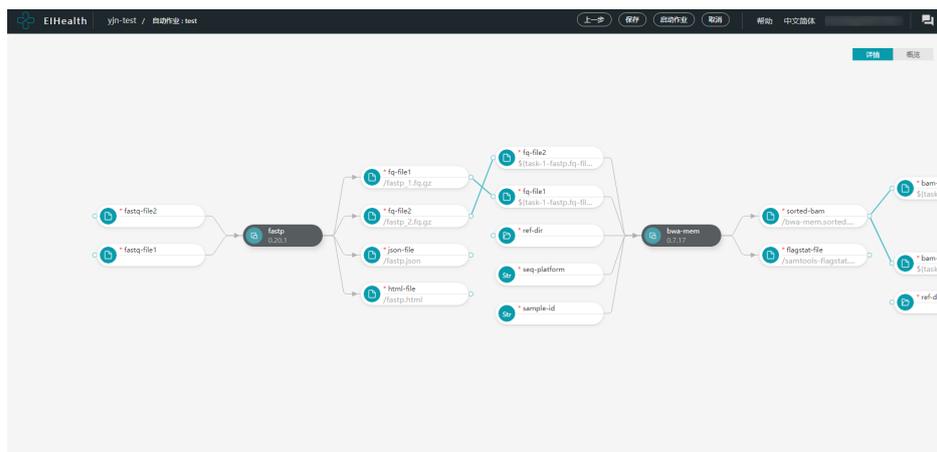
作业参数配置

- \* 作业名称: 自动生成
- 标签: 请输入标签名称  
最近常用标签: labels, label1, lab\_el-A
- 描述:
- 输出路径: 请选择
- 优先级: 低  高
- 计算节点标签:
- 超时时间: 1440 分钟
- 加速类型: 无 | **IO加速** | 本地盘加速
- 自动调度:

下一步: 设置流程参数 | 取消

6. 进入流程设计器页面。单击输入参数  图标，设置输入数据。

图 2-180 流程设计器



- 单击应用图标，弹出编辑图标，单击  按钮，设置应用参数。  
在“高级参数”下面可以设置加速类型。如果作业加速类型选择“无”，task加速类型可以任意选择。如果作业加速类型选择IO加速或者本地盘加速，则此处task加速类型无法设置。
- 参数确认无误后，单击页面上方“启动作业”按钮，可以直接基于该流程创建自动分析作业。

## 查看作业运行结果

自动作业运行完成后，在自动作业列表中单击数据表名称，跳转至对应的数据表中。在设置的状态更新列中，单击执行结果跳转至对应的分析作业，查看作业运行结果详情。

图 2-181 查看作业运行结果

| 作业名称     | 状态 | 数据表             | 创建时间                | 操作           |
|----------|----|-----------------|---------------------|--------------|
| test-bug | 停止 | test-bug        | 2022/06/20 11:38:12 | 启动   编辑   删除 |
| test123  | 失败 | long-double-150 | 2022/06/15 16:44:34 | 启动   编辑   删除 |
| sd0d     | 失败 | test-quota      | 2022/06/14 14:27:49 | 启动   编辑   删除 |
| 222      | 停止 | a0613           | 2022/06/14 17:47:06 | 启动   编辑   删除 |
| test-out | 停止 | string-out      | 2022/06/13 14:42:25 | 启动   编辑   删除 |
| aaadsfud | 失败 | alltype         | 2022/06/13 10:49:58 | 启动   编辑   删除 |
| test-111 | 停止 | string          | 2022/06/10 16:12:26 | 启动   编辑   删除 |
| -baKa-0  | 停止 | ced             | 2022/06/10 16:00:49 | 启动   编辑   删除 |
| iter6    | 停止 | iter6           | 2022/06/10 09:35:59 | 启动   编辑   删除 |
| test123  | 停止 | multi           | 2022/06/09 18:55:36 | 启动   编辑   删除 |

## 2.12.4 查看执行结果

分析作业的执行时间与环境资源类型、环境资源大小、处理数据大小等相关。您可以在“作业”页面查看执行结果或进行操作。提供了作业的执行状态的查看，并可对作业执行取消、删除、重试、克隆、导出操作。勾选多个作业，可以批量取消、批量删除、批量重试、批量导出。

作业执行状态如下所示。

- 等待中：作业创建成功，等待处理中。当作业在等待中，task状态全是unknown，时间条显示灰色。

- 运行中：作业创建成功，执行中。可单击作业名称，进入详情页查看运行详情。“运行中”的作业允许取消、克隆。
- 成功：作业运行成功。可单击作业名称，进入详情页查看运行详情。
- 失败：作业运行失败。可单击作业名称，进入详情页查看运行失败原因。
- 取消中：对“运行中”的作业，执行“取消”操作，取消运行。
- 已取消：取消运行成功。
- 强制取消：对于超时失败或者取消但未选择强制取消的场景，可以执行强制取消，强制作业。

## 查看执行状态

您可以通过单击作业名称，进入详情页面，查看详细的运行信息。包括运行状态、标签、描述、创建时间、完成时间、运行时间、计算节点标签、加速类型、优先级等。单击“概述”列 ▾ 按钮，在展开的信息栏中查看作业的输入&输出、节点参数、应用；在作业的子任务中可以查看日志、事件；在事件页签，可以查看实例的事件详情、下载YAML文件、查看监控。如果并发执行了多个作业，则会产生多个子任务。

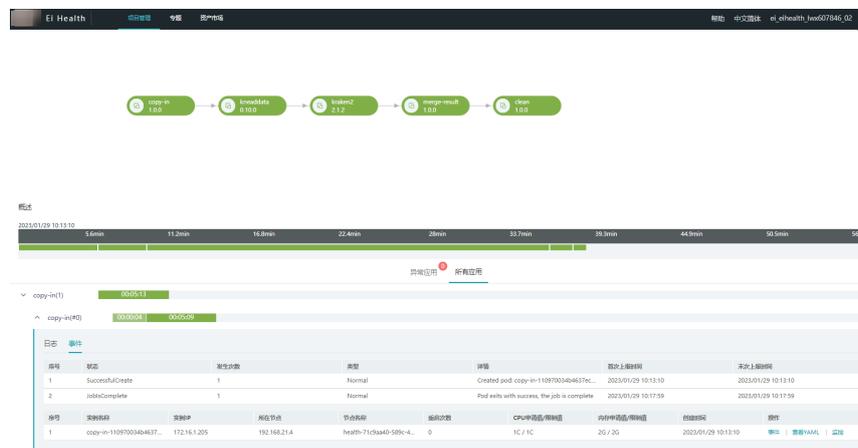
对于执行失败的作业，鼠标指向作业状态 ❌ 失败 📄，在弹出的提示框中可以查看“失败信息”和“失败原因”。同时，使用不同颜色提示执行状态，特别是对于由多个应用构成的分析作业，通过颜色方便地区分应用的执行状态。

- 绿色：运行成功。
- 红色：运行失败。
- 蓝色：等待运行。
- 灰色：被取消运行。
- 蓝色圆圈：运行中。

作业运行的时间，可以通过“概述”列的进度条进行查看。进度条中的颜色与应用状态颜色对应。单击进度条中的颜色块，可以展开并查看应用的运行日志，最多可以显示5000条日志。

分析作业创建后，可以通过“事件”查看容器执行该作业时的动作和状态，单击  图标，展开“事件”。

图 2-182 作业详情



## 📖 说明

如果作业显示运行正常，但实际作业中的某一个应用运行失败，请检查输入数据是否正常，并修改算法程序入口的main函数，保证运行结果有显式的返回值。

### • 故障说明

作业运行时，每个应用称之为一个任务（Task）。部分场景下，任务输入数据异常，实际作业运行失败，但界面显示运行正常。

作业中子任务没有返回正确的值，但是容器仍正常退出return 0，此时判定为子任务已正确执行。

实际上作业的子任务对应的就是K8S中的一个Pod，其返回状态就是Pod的phase映射，即容器以非0状态退出或者被系统终止会算作失败；容器return 0并且不再重启即算成功。详细内容介绍请参见[Pod的生命周期](#)。

### • 处理建议

所有的算法程序的入口main函数都显式的给出返回值，即正确执行则return 0。其他异常场景return其他数值或者抛出异常，并输出相关日志。

```
# python
def funcA():
    try:
        doSomething();
    except:
        log.print("An exception occurred, xxxx");
        raise 自定义异常;
def funcB():
    result = doSomething();
    return result;
if __name__ == '__main__':
    funcA();
    result = funcB();
    if result:
        sys.exit(1);
    else:
        sys.exit(0);
# c++
int main(){
    int result = doSomething();
    if(result != 0){
        return -1;
    }
    else{
        return 0;
    }
}
```

## 获取作业结果

作业运行成功后，可以看到每个应用的运行信息：输入输出，节点参数，应用以及日志等信息，最多可以显示5000条日志。

单击具体的输出参数可跳转到具体的文件位置。

图 2-183 获取作业结果



同时，项目支持导出作业的元数据信息，包括以下内容。

- 作业的基本配置信息，以yaml文件导出，包含作业名称，依赖的workflow名称/版本，基本的CPU/Memory配置，输入输出参数等信息。
- 作业的运行信息，包括每个任务的启动、停止时间、运行状态，日志链接等。

在作业对应的操作列中，单击“更多 > 导出”，导出单个作业的元数据信息。勾选多个作业，在上方单击“导出”，批量导出作业元数据信息。

图 2-184 导出作业元数据

| 作业名称                 | 版本 | 流程名称                    | 流程版本  | 状态  | 创建者              | 创建时间                | 完成时间                | 总耗时      | 操作    |
|----------------------|----|-------------------------|-------|-----|------------------|---------------------|---------------------|----------|-------|
| job-cancel-try       | -- | Variant-Calling-Base... | 1.0.0 | 成功  | ei_ehealth_jw... | 2022/06/20 14:57:25 | 2022/06/20 15:05:56 | 00:08:31 | 关闭 更多 |
| job-bingfa1gpu       | -- | test-bingfalog          | 1.2.3 | 成功  | ei_ehealth_jw... | 2022/06/20 14:55:34 | 2022/06/20 14:57:04 | 00:01:30 | 关闭 更多 |
| job-bingfa1gpu       | -- | test-bingfalog          | 1.2.3 | 成功  | ei_ehealth_jw... | 2022/06/20 14:46:05 | 2022/06/20 14:47:58 | 00:01:53 | 关闭 更多 |
| job-long-log         | g  | test-log                | 1.0.2 | 成功  | ei_ehealth_jw... | 2022/06/15 14:46:01 | 2022/06/15 14:51:02 | 00:03:01 | 关闭 更多 |
| job-cancel-try       | -- | Variant-Calling-Base... | 1.0.0 | 成功  | ei_ehealth_jw... | 2022/06/15 14:46:28 | 2022/06/15 17:31:01 | 02:24:33 | 关闭 更多 |
| job-cancel-try       | -- | Variant-Calling-Base... | 1.0.0 | 已完成 | ei_ehealth_jw... | 2022/06/15 10:35:27 | 2022/06/15 10:37:35 | 00:02:08 | 关闭 更多 |
| job-fail-try         | -- | Variant-Calling-Base... | 1.0.0 | 失败  | ei_ehealth_jw... | 2022/06/15 10:34:46 | 2022/06/15 14:37:37 | 04:02:51 | 关闭 更多 |
| job-force-cancel-try | -- | Variant-Calling-Base... | 1.0.0 | 成功  | ei_ehealth_jw... | 2022/06/15 10:11:41 | 2022/06/15 10:34:33 | 00:22:52 | 关闭 更多 |
| job-force-cancel     | -- | Variant-Calling-Base... | 1.0.0 | 已完成 | ei_ehealth_jw... | 2022/06/15 10:06:48 | 2022/06/15 10:10:18 | 00:03:30 | 关闭 更多 |
| job-cancel           | -- | Variant-Calling-Base... | 1.0.0 | 已完成 | ei_ehealth_jw... | 2022/06/15 09:48:29 | 2022/06/15 09:50:53 | 00:02:24 | 关闭 更多 |

## Task 失败状态条显示说明

每个Task 状态条显示成三段：Task添加重试失败总时间、当前pod等待时间、pod运行时间。

第一段表示所有重试失败的总时间，第二段表示当前pod等待时间，第三段表示当前pod运行时间。

如果所有重试pod都失败，则显示一个总的失败时间。如果没有重试，一次就执行成功，则无需显示失败时间，只显示等待时间+运行时间两段。

## 2.12.5 作业执行失败排查思路

作业执行失败，有如下场景：

- **场景1：**作业投递后处于运行中，运行过程正常，但是最后超时失败。
- **场景2：**作业投递后处于运行中，但是无日志打印，也没有任何符合预期的输出文件生成。
- **场景3：**作业投递后显示运行成功，但是根据日志分析作业有报错，也未能正确输出相关信息。
- **场景4：**参数配置的不合理导致的作业执行失败。
- **场景5：**并发投递多个作业后，部分作业运行失败，日志中显示IO读写相关错误。
- **场景6：**作业投递后运行失败，日志显示File name too long。
- **场景7：**同时复制多个文件，日志中显示cp: will not create hard link相关错误信息。

### 场景 1

作业投递后处于运行中，运行过程正常，但是最后超时失败。

|      |                                       |        |                     |
|------|---------------------------------------|--------|---------------------|
| 状态   | <span style="color: red;">❌</span> 失败 | 标签     | --                  |
| 创建时间 | 2022/09/22 10:01:34                   | 完成时间   | 2022/09/22 10:03:34 |
| 运行时间 | 00:02:00                              | 计算节点标签 | --                  |
| 加速类型 | 性能增强版 100GB                           | 描述     | --                  |
| 优先级  | 4                                     |        |                     |

workflow-0916 1.0



## 排查思路

判断是否是作业超时。

单击“状态”右侧的失败图标，查看失败原因，若为Graph timeOut for executin xxx，则可以确认为作业超时失败。

|      |                                       |                                                                         |                     |
|------|---------------------------------------|-------------------------------------------------------------------------|---------------------|
| 状态   | <span style="color: red;">❌</span> 失败 | 失败信息: Execution TimeOut                                                 | --                  |
| 创建时间 | 2022/12/06 20:52:59                   | 失败原因: Graph timeOut for execution(3124c5f8-33d4-45f1-b00f-e4367fc8ee66) | 2022/12/07 20:52:59 |
| 运行时间 | 01天 00:00:01                          |                                                                         | --                  |
| 加速类型 | --                                    |                                                                         | --                  |
| 优先级  | 9                                     |                                                                         | --                  |

## 解决方法

根据实际业务需要，在“创建作业”阶段，设置合适的超时时间，默认设置的超时时间为24小时（1440分钟），最大可设置为144000分钟，即作业运行至多100天。

图 2-185 作业设置

作业设置
✕

\* 作业名称

标签   
最近常用标签: labels label1 lab\_el-A

描述   
0/255

\* 流程名称

输出路径

优先级 低  高

计算节点标签

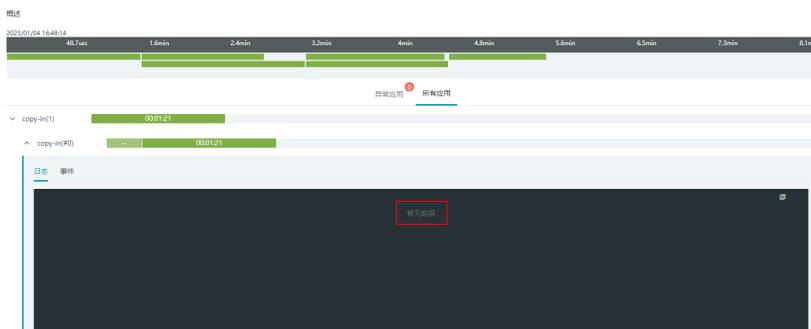
超时时间  分钟

加速类型

确定
取消

## 场景 2

作业投递后处于运行中，但是无日志打印，也没有任何符合预期的输出文件生成。



### 排查思路

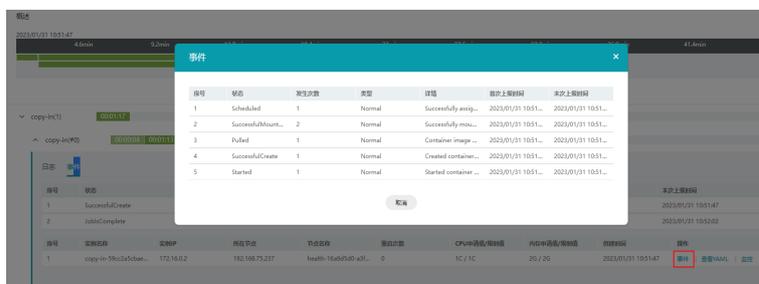
1. 首先需要用户自行确认一下投递的作业是否会在控制台打印日志，如果是有重定向日志输出到具体文件的话，此处无日志为正常现象。
2. 子任务的事件中，确认作业子任务的实例是否有正常创建。

图 2-186 子任务的事件



3. 查看实例的事件，查看实例是否有正常创建。

图 2-187 实例的事件



### 解决方法

- 若子任务未正常创建，请联系服务技术支持解决。
- 若子任务正常创建，但是实例未正常创建，可以通过事件信息分析，常见有以下问题场景和对应解决方案。
  - 0/4 nodes are available: XXX Insufficient cpu 或者 XXX Insufficient memory。该场景表示当前集群中无充足的计算资源，可以根据实际需要提前结束掉其他作业或notebook来释放资源，也可以进入系统资源页面购买新节点。
  - 0/4 nodes are available: XXX node(s) didn't match node selector。该场景表示当前集群中无计算资源满足标签要求，用户可以进入系统资源页面，选择节点，通过标签管理给节点添加标签。
  - 其他场景可以联系服务技术支持解决。

图 2-188 购买计算资源



图 2-189 设置计算节点标签



### 场景 3

作业投递后显示运行成功，但是根据日志分析作业有报错，也未能正确输出相关信息。

排查思路、解决方法请参考[查看执行结果](#)章节的说明。

**说明：**  
如果作业显示运行正常，但实际作业中的某一个应用运行失败，请检查输入数据是否正确，并修改算法程序入口的main函数，保证运行结果有显式的返回值。

- 故障说明**  
作业运行时，每个应用称之为一个任务（Task），部分场景下，任务输入数据异常，实际作业运行失败，但界面显示运行正常。作业中子任务没有返回正确的值，但是容器仍正常退出return 0，此时判定为子任务已正确执行。实际上作业的子任务对应的就是K8S中的一个Pod，其返回状态就是Pod的phase映射，即容器以非0状态退出或者被系统终止会算做失败；容器return 0并且不再重启即算成功。详细内容介绍请参见Pod的生命周期。
- 处理建议**  
所有的算法程序的入口main函数都要显式的给出返回值，即正确执行则return 0，其他异常场景return其他数值或者抛出异常，并输出相关日志。

```
# python
def funcA():
    try:
        doSomething();
    except:
        log.print("An exception occurred, xxxxxx");
        raise 自定义异常;
def funcB():
    result = doSomething();
    return result;
if __name__ == '__main__':
    funcA();
    result = funcB();
    if result:
        sys.exit(1);
    else:
        sys.exit(0);
```

### 场景 4

参数配置的不合理导致的作业执行失败。

解决方法

1. 选择运行失败的分析作业，单击操作列“更多>重试”。
2. 在弹出的提示框中选择“更改参数”。

图 2-190 更改参数



3. 在作业配置页面，修改相关参数后，单击页面右上角的“重试作业”。

图 2-191 重试作业



## 场景 5

并发投递多个作业后，部分作业运行失败，日志中显示IO读写相关错误。

图 2-192 运行出错的应用日志



### 排查思路

检查同一时刻投递作业的输出路径是否存在重复。若存在重复，则很有可能是并发写入同一个文件导致的异常，若不存在请联系服务技术支持解决。

### 解决方案

平台提供了作业级输出路径，流程级输出路径，子任务级输出路径用于做不同层级的文件隔离。如下图所示，如果三者填写的路径相同，请修改输出路径后重试，如果填写的不同，请联系服务技术支持解决。当三者均不填写时，平台会自动生成随机路径。

图 2-193 作业级输出路径

**新建作业**

\* 作业名称: job-9772

标签: 请输入标签名称  
最近常用标签: labels label1 lab\_el-A

描述: 0/255

\* 流程名称: 请选择想要执行的流程

**输出路径** ?

优先级: 若不指定, 系统将按job名称-UUID自动生成job的根路径。 高

计算节点标签 ?

超时时间: 1440 分钟

加速类型 ? 无 IO加速 本地盘加速

自动调度:

确定 取消

图 2-194 流程级输出路径

**流程设置**

\* 流程名称: 流程名称

\* 版本: 流程版本

标签: 请输入标签名称  
最近常用标签: 客户1 smallMolecularC peptide-docking

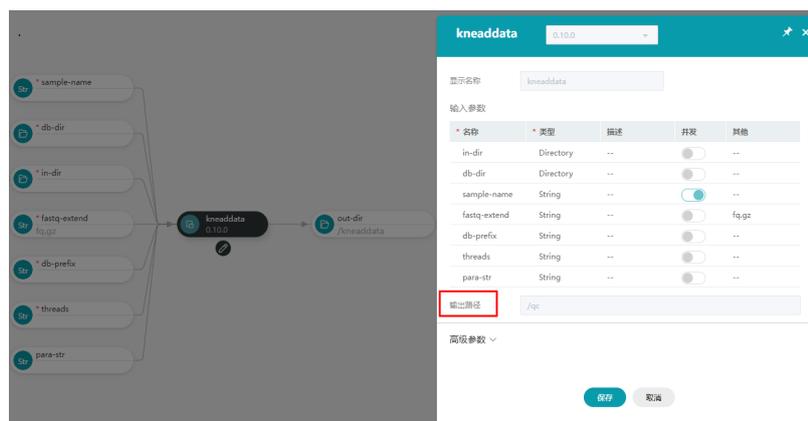
短描述: 0/128

描述: 0/65535

超时时间: 1440 mins

**输出路径**

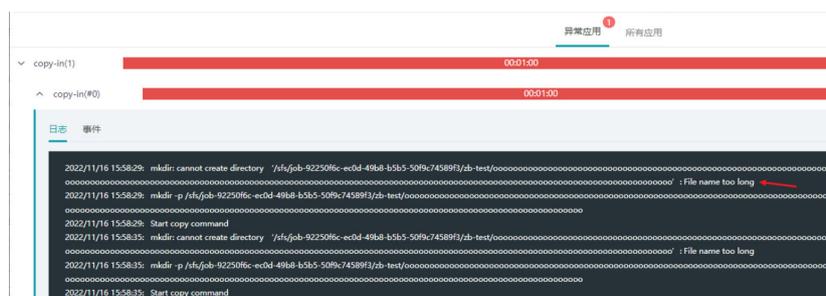
图 2-195 子任务级输出路径



## 场景 6

作业投递后运行失败，日志显示File name too long。

图 2-196 运行出错的应用日志



### 排查思路

检查输入文件的路径单级名称长度是否超过255个字符。若单级名称长度未超过255个字符请联系技术支持。

### 解决方案

1. 单级目录名称长度超过255个字符。
  - 不使用SFS或者EVS加速。
  - 若必须进行IO加速，则更改输入文件的路径为单级目录名称不超过255字符的路径。
2. 若单级名称长度未超过255个字符请联系技术支持。

## 场景 7

同时复制多个文件，日志中显示cp: will not create hard link相关错误信息

图 2-197 输入输出参数信息

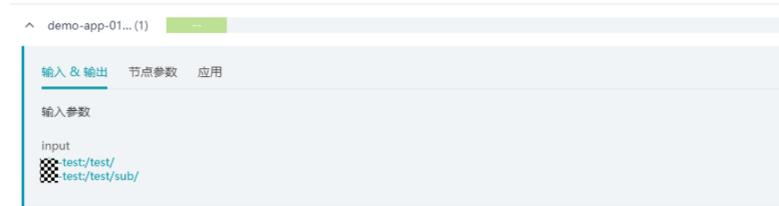


图 2-198 应用信息



图 2-199 失败日志信息



### 排查思路

检查作业是否存在文件或目录类型的输入参数，并且未开启并发，同时改输入参数还填入了多个值，并且路径存在包含关系，如上图所示。

假定镜像命令为 `cp -rf ${input} ${output}`，变量替换后实际执行命令为 `cp -rf /test/sub /test/output`，此时会触发 `cp: will not create hard link` 的错误。若不符合上述场景，请联系技术支持。

### 解决方案

1. 首先排查业务场景是否需要涉及并发，若涉及，请前往应用详情页面开启并发选项即可，则不会出现本问题场景。

图 2-200 开启并发



2. 若不涉及并发，则优先排查镜像启动命令设置是否合理、是否确实需要复制路径存在包含关系的多个文件或目录。若不存在相关场景，根据实际业务需求，修改镜像启动命令即可。
3. 若确实存在路径需要包含多个文件或目录，可参考如下方案解决。
  - a. 以下图为例，首先调整应用的输入参数为 `input-a` 和 `input-b`，分别表示两个路径存在包含关系的目录（`/test/sub` 和 `/test`）。



- b. 然后修改应用的镜像启动命令如下即可。

```
cp -rf ${input-a} ${output};
cp -rf ${input-b} ${output};

# 变量替换后实际执行命令会是
cp -rf /test/sub /output;
cp -rf /test /output;
```

## 2.13 Nextflow

### 2.13.1 简介

Nextflow是专门针对生物信息流程搭建而开发的一种框架语言，它能够很好的管理生信流程，并且将其与Conda、Docker、Singularity结合起来使用，可以很好的将流程在不同平台之间进行迁移，并且能够保证结果的可重复性。Nextflow最大的优点是它是基于数据流的程序模型，因此不用自己去写复杂的并行化的程序，只需要关注“数据流”即可。流程中的每个process可以通过配置文件快速的定制并行执行的样本数目，大大节约了生信人员搭建组学分析流程的时间。

医疗智能体（EIHealth）平台集成了Nextflow框架，支持用户直接在医疗智能体（EIHealth）平台安装Nextflow。

### 2.13.2 安装/卸载 Nextflow

#### 安装 Nextflow

Nextflow需要用户自己进行安装，安装的权限只有管理员拥有。

1. 在平台右上角用户名中选择“系统设置”。
2. 在“Nextflow配置”模块，单击“安装”。

图 2-201 安装 Nextflow



3. 选择安装方式。目前支持GitHub获取和上传安装包两种方式。
  - GitHub获取：可以选择版本，目前支持选择的安装包版本v22.10.6, v22.10.0, v22.04.0。
  - 上传安装包：用户直接上传一个Nextflow的安装包。如果选择此方式，您可以单击“点击此处”进入nextflow release进行下载安装包。

图 2-202 选择安装方式



4. 单击“开始安装”。

## 卸载 Nextflow

1. 在平台右上角用户名中选择“系统设置”。
2. 在“Nextflow配置”模块，单击“卸载”。

图 2-203 卸载 Nextflow



3. 在弹出的对话框中输入“yes”，单击“确定”。



## 2.13.3 新建流程

1. 选择“工具>Nextflow”，单击“新建流程”。
2. 填写流程基本信息。
  - 名称：设置流程名称。名称长度为1-56，只允许出现中划线、下划线、字母和数字。

- 标签：选择流程标签，最多可添加5个。
  - 描述：设置流程相关描述。
3. 单击“下一步”，添加流程文件。  
单击“流程文件”后面的“添加”，添加流程文件。  
如果流程文件上传的是zip包，则可以选择添加“主文件”。主文件必须为.nf，为压缩文件夹中的文件完整路径。例如：main.nf。若不填写，默认为main.nf。

#### 📖 说明

- 支持上传的文件格式为nf、zip，大小不能超过10M。zip包文件个数规格为1024个，单个文件解压后限制，不超过10M。
- 如果上传的流程中包含镜像，单击镜像管理跳转至本项目的镜像管理列表，进行镜像上传，具体可参考[导入或上传镜像](#)。

图 2-204 添加流程文件



4. 添加完成后，单击“下一步”，配置流程参数。

添加参数方式支持两种，一种为单击  添加参数 进行添加，一种为单击

 上传参数 进行添加。如果参数比较多，建议使用“上传参数”方式。

- 上传参数方式：支持json或yaml格式，文件大小不能超过10M。模板可以通过单击  下载模板 进行下载。如果上传参数有重名的参数，则只取最后一个参数的值。
- 添加参数方式：需手动配置相关参数的取值。
  - 参数名称：设置参数名称，此参数为必填项。参数名不能相同。
  - 类型：参数类型，此参数为必填项。取值范围：Other、File、Directory。如果类型选择Other时，可以单击值旁边的  按钮进行输入。如果选择File或者Directory类型，可以单击值旁边的  按钮，选择文件或者文件夹。

- 值：根据参数类型，填写或者选择相关取值。
- 必传：参数是否必传，此参数为非必填项。如果开启“必传”，则“值”为必填项。
- 描述：参数描述，此参数为非必填项。
- 操作：可以单击操作列“删除”，删除对应的参数。

图 2-205 配置流程参数



5. 参数配置完成后，单击“保存”。

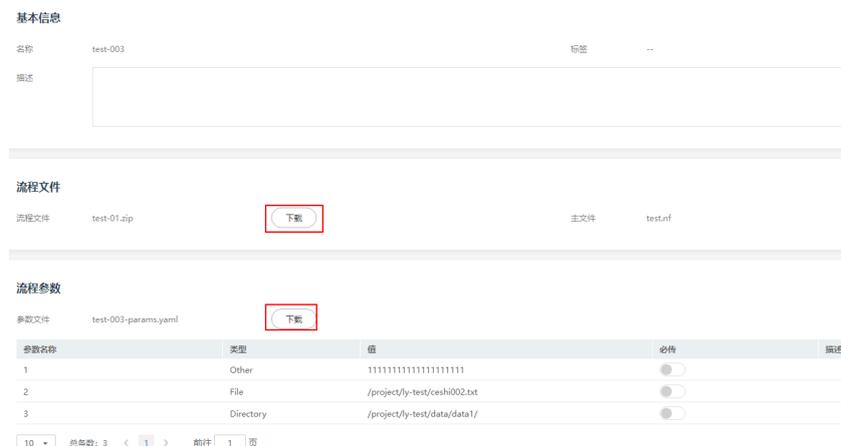
流程创建成功后，单击“保存并启动作业”，或者在流程管理页面，单击流程列表操作列的“启动作业”进入作业设置页面，配置作业的相关信息，具体可参考[新建作业](#)。

## 2.13.4 管理流程

流程创建成功后，您可以对流程进行编辑，删除操作。

- 编辑  
选择已创建成功的流程，单击操作列的“编辑”，进入流程编辑页面。除了流程名称不能更改，您可以修改已有流程的其他参数配置项。
- 删除  
通过单击操作列“删除”，可删除对应的流程。
- 下载流程文件  
单击流程名称，进入流程详情页面，单击流程文件后的“下载”，可下载该流程文件。
- 下载流程参数  
单击流程名称，进入流程详情页面，单击流程参数后的“下载”，可下载该流程设置的参数文件。

图 2-206 下载流程文件/参数

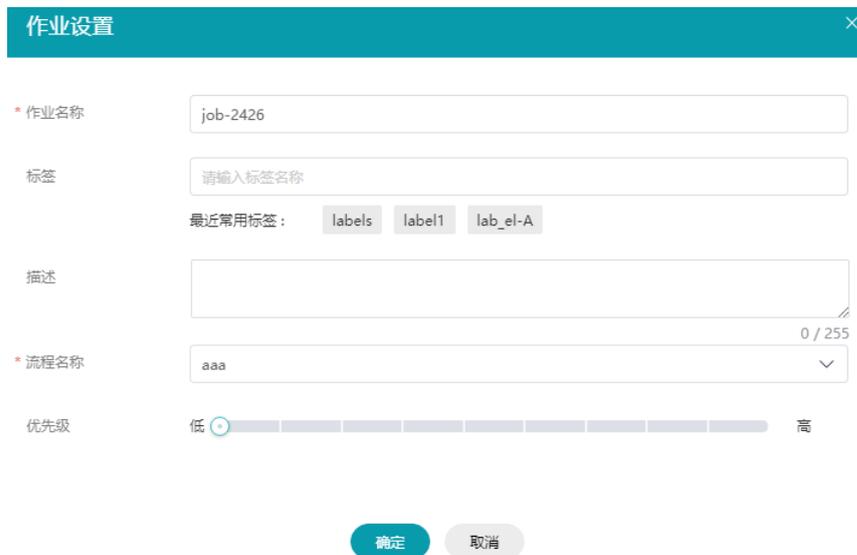


## 2.13.5 新建作业

Nextflow作业依托于流程运行，需要您先[新建流程](#)，再基于流程创建作业。

1. 选择“作业>分析作业>Nextflow”，单击“新建作业”。
2. 设置作业参数。
  - 作业名称：设置作业名称。名称长度为1-63，只允许出现中划线、字母和数字。
  - 标签：设置作业标签。最多可以设置5个。
  - 描述：作业描述。
  - 流程名称：选择已创建的流程，如果还未创建，可参考[新建流程](#)进行创建。
  - 优先级：设置作业优先级。在资源充足的情况，高优先级先于低优先级。在满足条件的资源规格情况下，高优先级的先于低优先级投递。如果资源不满足，则优先处理资源满足的作业。例如，资源有2u2g，低优先级的需要1.8u1.8g，高优先级的需要2.2u2.2g，则低优先级的会先投递。但是两者都是1.8u1.8g，高优先级的先投递。

图 2-207 新建作业



3. 单击“确定”，进入启动作业页面。

4. （可选）在启动作业页面，可以通过 ，添加作业相关参数。

上传参数的模板可以通过单击  进行下载。

如果流程设置了参数，作业上传的参数将会覆盖流程设置的参数。

5. 单击“启动作业”。

## 2.13.6 管理作业

### 作业状态

- 运行中：作业创建成功，执行中。可单击作业名称，进入详情页查看运行详情。“运行中”的作业允许取消、克隆。
- 失败：作业运行失败。可单击作业名称，进入详情页查看运行失败原因。失败的作业允许重试、删除、克隆。
- 成功：作业运行成功。可单击作业名称，进入详情页查看运行详情。成功的作业允许删除、克隆。
- 已提交：作业创建成功，但未运行。已提交的作业允许取消、克隆。
- 已取消：取消已提交或者运行中的作业。已取消的作业允许重试、删除、克隆。

### 查看作业详情

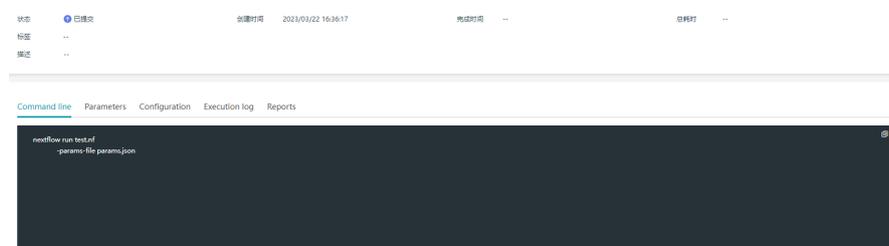
您可以通过单击作业名称，进入详情页面，查看详细的运行信息。包括状态、标签、描述、创建时间、完成时间、总耗时等。

对于执行失败的作业，鼠标点击作业状态  失败 ，在弹出的提示框中可以查看“失败信息”和“失败原因”。

### 作业相关信息

在详情页可以查看作业执行命令、配置参数、流程配置文件、执行日志、报告信息。

图 2-208 作业相关信息



### Tasks相关信息

在详情页可以查看每个Tasks的相关信息。单击每个task可以打开Task详情，查看对应的processname和执行日志。

图 2-209 Tasks 相关信息

| task_id | process        | tag | hash      | status | container | native_id          | submit              | complete            | duration | realtime | %cpu | %mem |
|---------|----------------|-----|-----------|--------|-----------|--------------------|---------------------|---------------------|----------|----------|------|------|
| 1       | splitLetters   |     | c5/1010c7 | FAILED |           | nf-c51010c7acc...  | 2023/03/22 16:36:47 | 2023/03/22 16:36:57 | 10s      | 413ms    | 13.7 | 0    |
| 2       | convertToUpper |     | c5/cff4df | FAILED |           | nf-c5cff4df#fc2... | 2023/03/22 16:37:02 | --                  | --       | --       | --   | --   |

## 管理作业

- 克隆  
对于已创建的作业，可以单击操作列“克隆”，进行克隆操作。
- 删除  
对于“失败”、“成功”、“已取消”的作业均可删除。单击对应作业操作列“更多>删除”，删除对应的作业。
- 取消  
对于“运行中”和“已提交”的作业，可以执行取消操作。单击对应作业操作列“更多>取消”，取消对应的作业。
- 重试  
对于状态为“失败”或者“已取消”的作业，可以选择重试。单击对应作业操作列“更多>重试”，根据提示对话框，选择“立即重试”或者“更改参数”。  
如果选择“更改参数”，系统会跳转至作业参数配置页面，进行修改后，单击“重试作业”。
- 搜索
  - 普通搜索  
在作业管理页面，右上角搜索框中输入作业名称，单击  按钮进行搜索。

图 2-210 搜索作业



- 高级搜索  
Nextflow作业支持根据流程名称、标签、状态、创建者、创建时间、完成时间进行搜索。在作业管理右上角，选择“高级搜索”，输入搜索条件后，单击“搜索”即可查找出相关信息。

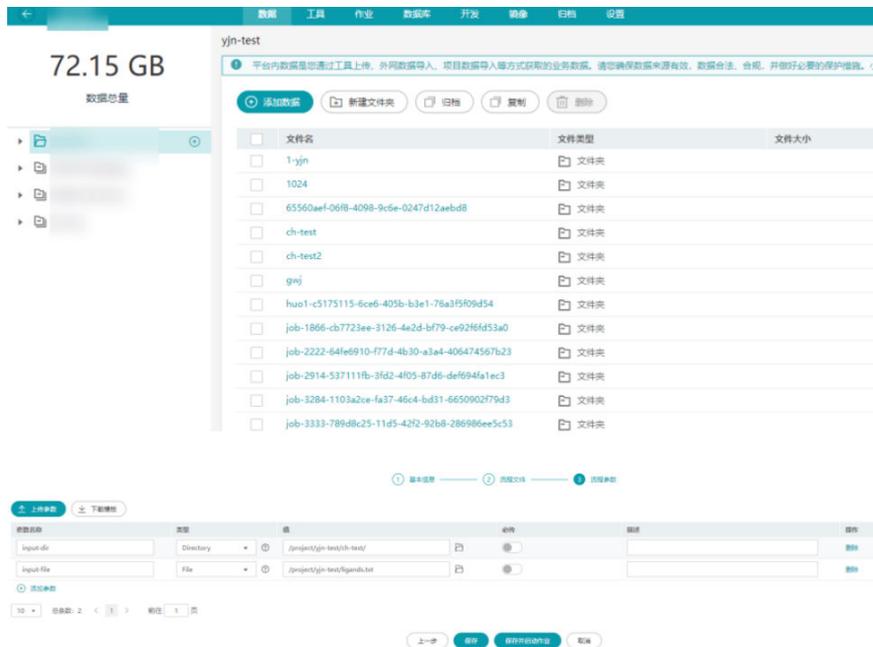
图 2-211 Nextflow 作业高级搜索



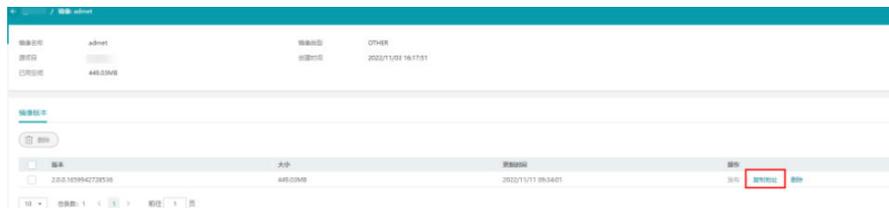
## 2.13.7 约束限制

当前Nextflow相关特性仅为公测版本，部分功能暂不支持，因此存在如下限制约束。

- **流程使用的镜像和数据需要按照医疗智能体（EIHealth）平台的规范进行上传**  
推荐通过eihealth-toolkit进行上传。eihealth-toolkit介绍详见[什么是医疗智能体eihealth-toolkit](#)。数据上传可参考[上传数据](#)。  
数据上传成功后可以在数据详情页面查询，创建Nextflow流程时可以通过设置参数指定对应数据。



镜像上传可参考[上传镜像](#)。镜像上传成功后可以在镜像详情页面查询，复制其镜像地址填入Nextflow脚本中的container字段即可。



```
process test { container 'swr.${regionID}.myhuaweicloud.com/eihealth/admet:2.0.0' }
```

- **暂不支持用户设置部分Nextflow配置项**

对于dag, timeline, report, trace配置，医疗智能体（EIHealth）平台会默认开启并且指定相关路径，会覆盖用户侧配置，可以在作业详情页面查询和下载相关信息。相关信息可参考<https://www.nextflow.io/docs/latest/tracing.html>



对于pod配置，暂不支持用户侧设置。若用户侧设置的情况下可能会导致对应Process执行失败。相关信息可参考<https://www.nextflow.io/docs/latest/process.html#pod>

The `pod` directive allows the definition of the following options:

|                                                                        |                                                                                                                                                                                                                         |
|------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <code>label: &lt;K&gt;, value: &lt;V&gt;</code>                        | Defines a pod label with key <code>K</code> and value <code>V</code> .                                                                                                                                                  |
| <code>annotation: &lt;K&gt;, value: &lt;V&gt;</code>                   | Defines a pod annotation with key <code>K</code> and value <code>V</code> .                                                                                                                                             |
| <code>env: &lt;E&gt;, value: &lt;V&gt;</code>                          | Defines an environment variable with name <code>E</code> and whose value is given by <code>V</code> .                                                                                                                   |
| <code>env: &lt;E&gt;, fieldPath: &lt;V&gt;</code>                      | Defines an environment variable with name <code>E</code> and whose value is given by the field <code>V</code> .                                                                                                         |
| <code>env: &lt;E&gt;, config: &lt;C/K&gt;</code>                       | Defines an environment variable with name <code>E</code> and whose value is given by the config <code>C</code> with key <code>K</code> .                                                                                |
| <code>env: &lt;E&gt;, secret: &lt;S/K&gt;</code>                       | Defines an environment variable with name <code>E</code> and whose value is given by the secret <code>S</code> with key <code>K</code> .                                                                                |
| <code>config: &lt;C/K&gt;, mountPath: &lt;/absolute/path&gt;</code>    | Mounts a <code>ConfigMap</code> with name <code>C</code> with key <code>K</code> to the path <code>/absolute/path</code> . When the key component is omitted the path is interpreted as <code>"/absolute/path"</code> . |
| <code>csi: &lt;V&gt;, mountPath: &lt;/absolute/path&gt;</code>         | Mounts a <code>CSI ephemeral volume</code> with config <code>V</code> to the path <code>/absolute/path</code> .                                                                                                         |
| <code>emptyDir: &lt;V&gt;, mountPath: &lt;/absolute/path&gt;</code>    | Mounts an <code>emptyDir</code> with configuration <code>V</code> to the path <code>/absolute/path</code> .                                                                                                             |
| <code>secret: &lt;S/K&gt;, mountPath: &lt;/absolute/path&gt;</code>    | Mounts a <code>Secret</code> with name <code>S</code> with key <code>K</code> to the path <code>/absolute/path</code> . When the key component is omitted the path is interpreted as <code>"/absolute/path"</code> .    |
| <code>volumeClaim: &lt;V&gt;, mountPath: &lt;/absolute/path&gt;</code> | Mounts a <code>Persistent volume claim</code> with name <code>V</code> to the specified path <code>lc</code> parameter to mount a directory inside the referenced volume instead of <code>"/absolute/path"</code> .     |
| <code>imagePullPolicy: &lt;V&gt;</code>                                | Specifies the strategy to be used to pull the container image e.g. <code>imageIfNotPresent</code> .                                                                                                                     |
| <code>imagePullSecret: &lt;V&gt;</code>                                | Specifies the secret name to access a private container image registry. <code>S</code> .                                                                                                                                |

对于publishDir配置，用户侧设置时需要使用一定的技巧。目前测试结果表明下面两种写法是可行的，供参考。

写法1:

```
process A {
  ... publishDir "/project/guancha/test_publishdir/", mode:"copy"

  ... input:
  ... path outdir
  ... path infile

  ... output:
  ... path "test.txt"

  ... script:
  ... """
  ... head -n 10 $infile > test.txt
  ... """
}

workflow {
  ... A(
  ... | ... params.outdir,
  ... | ... params.infile
  ... )
}
```

写法2:

```
process A {
  publishDir params.outdir, mode:"copy"

  input:
  path outdir
  path infile

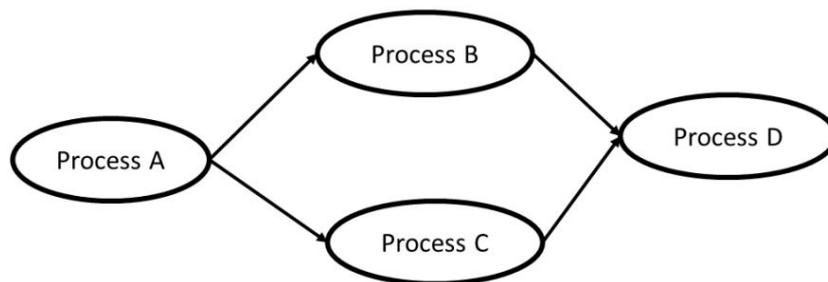
  output:
  path "test.txt"

  script:
  """
  head -n 10 $infile > test.txt
  """
}

workflow {
  A(
    params.outdir,
    params.infile
  )
}
```

### 正向用例

如下所示为一个不带业务属性（仅执行基本数学运算）的正向用例，包含了输入输出，串并联场景，适用面较广。



```
process PROCESS_A {
  input:
  val x
  output:
  val y
  exec:
  y = x*10
}
process PROCESS_B {
  input:
  val x
  output:
  val y
}
```

```

when:
x <= 50

exec:
y = x*10
}
process PROCESS_C {
input:
val x
output:
val y
when:
x > 50
exec:
y = x*10
}
process PROCESS_D {
debug true
input:
val x
output:
stdout
script:
"""
echo $x
"""
}
workflow {
Channel
.of(1..10)
| PROCESS_A
| PROCESS_B & PROCESS_C
| mix
| PROCESS_D
}

```

基于Nextflow官方提供first script，预期输出会打印Hello world的日志。

<https://www.nextflow.io/docs/latest/getstarted.html#your-first-script>

```

#!/usr/bin/env nextflow
nextflow.enable.dsl=2
params.str = 'Hello world!'
process splitLetters {
// 替换为正确镜像
container 'swr.XXX.myhuaweicloud.com/XXX/admet:2.0.0'
output:
path 'chunk_*'
"""
printf '${params.str}' | split -b 6 - chunk_
"""
}
process convertToUpper {
input:
path x
output:
stdout
"""
cat $x | tr '[a-z]' '[A-Z]'
"""
}
workflow {
splitLetters | flatten | convertToUpper | view { it.trim() }
}

```

## 2.13.8 作业执行失败排查思路

### 启动作业后作业一直处于已提交的状态

#### 问题现象

作业投递后一直处于已提交的状态。

#### 问题排查和解决方案

1. 查看execution log，若execution log为空，请提交工单或联系服务技术支持。
2. execution log 提示K8S pod can 't be scheduled，请根据日志信息扩容计算资源。
3. execution log提示can't lock file，重试或者克隆作业即可。
4. execution log提示java.text.ParseException: Unparseable date，请提交工单或联系服务技术支持。
5. execution log提示其它因脚本语法的问题，请根据报错信息进行修改，若无法解决请提交工单或联系服务技术支持。

### 作业状态已为完成态，task 仍在运行中

#### 问题现象

作业的状态已为完成态（失败、成功、取消），但是仍有部分task处于运行中。

#### 解决方案

无需特殊处理，以作业状态为准，查看异常的task日志。

### 重试作业或者删除作业出现内部异常

#### 问题现象

启动作业立即取消作业，进行重试或者删除作业出现内部异常。

#### 问题原因

投递作业属于异步动作，立即取消可能存在作业未完成投递，Nextflow对于不存在的作业做了静默处理。

#### 解决方案

1. 避免启动作业后立即取消。
2. 克隆作业即可。

### process 使用 errorStrategy 'ignore'时整个作业不会结束

#### 问题现象

process中使用了errorStrategy 'ignore'，当该process因为某种原因失败后，整个作业的状态一直处于运行中。

#### 问题原因

失败的task被忽略，但是后续依赖该process的task会被挂起。

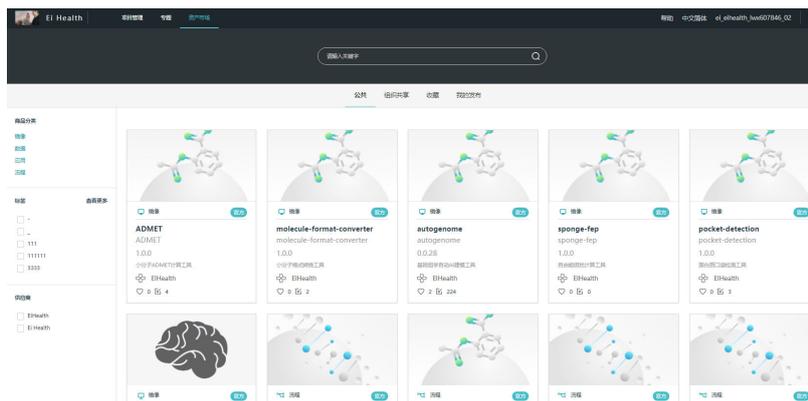
#### 解决方案

不建议使用errorStrategy 'ignore'。

## 2.14 资产市场

### 2.14.1 资产市场简介

资产市场是在EIHealth平台的基础上构建的开发者生态社区，提供镜像、数据、应用和流程的内容共享。提供安全、开放的资产共享和使用，加速基因、药物、临床应用的开发与落地，保障开发生态链上各参与方高效地实现各自的商业价值。

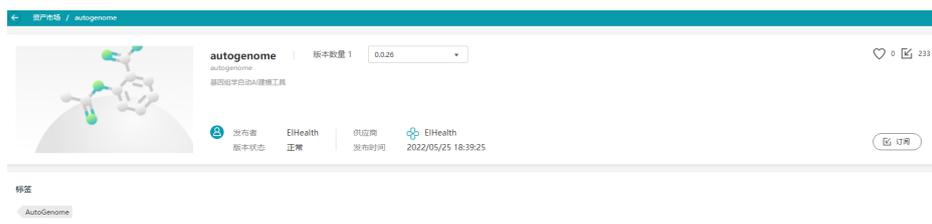


### 订阅资产

资产市场提供官方发布的镜像、数据、应用、流程资产。您可以查看各类资产的详细介绍和使用方法，并订阅资产至项目中使用。

1. 单击所需的资产名称，查看资产的详细信息。
2. 单击界面右侧“订阅”图标，订阅资产。

图 2-212 订阅资产

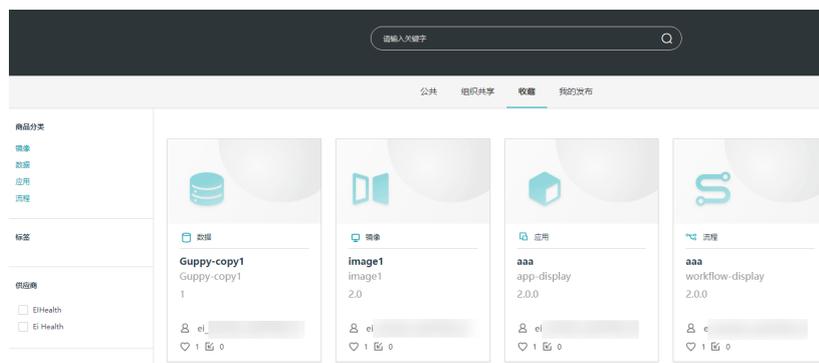


3. 订阅的资产将显示在您的项目中。对于不使用的资产，您可以通过“删除”操作，取消资产订阅。
  - 订阅的流程、应用类资产，将显示在“项目管理 > 工具”页面的列表中，以“[资产市场]”进行标记。
  - 订阅的镜像类资产，将显示在“项目管理 > 镜像”页面的列表中，以“[资产市场]”进行标记。
  - 订阅的数据类资产，将显示在“项目管理 > 数据”页面。

## 收藏资产

资产市场提供了官方发布的镜像、数据、应用、流程资产，您可以进入资产页，单击右上角的  图标，收藏资产，将与业务相关的资产放置收藏列表中，如下图所示，方便后期的查找和使用。对于已收藏的资产，可单击心形图标，取消收藏。

图 2-213 收藏资产



## 公共资产列表

表 2-23 资产列表

| 分类 | 资产名称                         | 说明                                                                                |
|----|------------------------------|-----------------------------------------------------------------------------------|
| 镜像 | image-stitching              | 针对TB级3D鼠脑稀疏标记成像数据的全自动拼接软件，减少数据分析50%的错误率，节省20%的时间，助力脑科学研究的效率提升。                    |
|    | AutoGenome                   | AutoGenome为Notebook镜像，是一个利用AutoML等技术帮助科研工作者在基因组学数据上端到端实现深度学习网络搜索，训练，评估，预测和解释的工具包。 |
| 数据 | 人基因组数据                       | GRch38-reference数据集为人类基因参考基因组，广泛用于人类基因组分析中，如WGS、callvariants等。数据集总大小约 13GB。       |
|    | NGS大数据集                      | NA24385-raw数据集为NGS流程测试数据集，作为该流程的原始输入。数据集总大小约 186.2GB。                             |
|    | NGS小数据集                      | NA12878-small数据集为NGS流程测试数据集，作为该流程的原始输入。数据集总大小约 216MB。                             |
|    | docking summary<br>测试数据      | 配体文件：小分子化合物SMILES结构式文件。<br>受体文件：蛋白3D结构PDB文件。                                      |
|    | RNA-Seq测试数据<br>及参考基因组数据<br>集 | RNA-Seq-Dataset数据集包含RNA-Seq分析流程的测试数据（fastq）和流程包含软件STAR依赖的参考基因组。                   |

| 分类 | 资产名称                           | 说明                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
|----|--------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 流程 | Variant Calling Based On NGS   | <p>二代基因组测序即Next Generation Sequencing (NGS)是一种基于边合成边测序的方式。NGS在保持了测序高准确度的同时,大幅的提高了测序速度。目前NGS已经普遍的应用于全基因组测序、全外显子测序、转录组测序、表观遗传学等领域。</p> <p>该流程以NGS得到的fastq作为输入,通过质控,比对,得到比对后的bam文件,及对fastq和bam文件的质控报告。</p>                                                                                                                                                                                                                                                         |
|    | Docking Summary                | <p>对一组小分子化合物配体和一组蛋白受体进行分子对接,汇总分子对接结果,用于可视化展示。</p> <p>该流程主要完成的功能包括:整合分子对接结果,生成结合能矩阵、整合受体与分子对接产生的配体构象,进行可视化展示、对配体分子进行注释,包括:DrugBank编号、分类、化学式、XLOGP3、TPSA、靶点、Csp3比例、分子量、可旋转键数目。</p>                                                                                                                                                                                                                                                                                |
|    | RNA-Seq Analysis Based on STAR | <p>二代基因组测序即Next Generation Sequencing (NGS)是一种基于边合成边测序的方式。NGS在保持了测序高准确度的同时,大幅的提高了测序速度。目前NGS已经普遍的应用于全基因组测序、全外显子测序、转录组测序和表观遗传学等领域。</p> <p>该流程以NGS得到的fastq作为输入,通过质控,比对等步骤,输出针对fastq的qc报告,输出STAR比对得到的bam文件。</p>                                                                                                                                                                                                                                                     |
|    | MetaGenome Kraken2 pipeline    | <p>宏基因组 (Metagenome) (也称微生物环境基因组 Microbial Environmental Genome, 或元基因组) 是由 Handelsman 等1998年提出的新名词,其定义为 “the genomes of the total microbiota found in nature”, 即环境中全部微小生物遗传物质的总和。它包含了可培养的和未可培养的微生物的基因,目前主要指环境样品中的细菌和真菌的基因组总和。宏基因组学(或元基因组学, metagenomics)是一种以环境样品中的微生物群体基因组为研究对象,以功能基因筛选和/或测序分析为研究手段,以微生物多样性、种群结构、进化关系、功能活性、相互协作关系及与环境之间的关系为研究目的的新的微生物研究方法。一般包括从环境样品中提取基因组DNA,进行高通量测序分析,或克隆DNA到合适的载体,导入宿主菌体,筛选目的转化子等工作。</p> <p>该流程主要基于Kraken2构建,跟进数据库进行物种注释。</p> |

| 分类 | 资产名称                                         | 说明                                                                                                                                                                                                                                                            |
|----|----------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|    | RNA Cufflinks transcriptome analysis process | 二代基因组测序即Next Generation Sequencing (NGS)是一种基于边合成边测序的方式。NGS在保持了测序高准确度的同时，大幅的提高了测序速度。目前NGS已经普遍的应用于全基因组测序、全外显子测序、转录组测序和表观遗传学等领域。<br>该流程以NGS得到的SRA文件作为输入，通过拆分 reads、fastqc质控、tophat2比对，然后 Cufflinks 利用Tophat比对的结果（alignments）来组装转录本，估计这些转录本的丰度，并且检测样本间的差异表达及可变剪接。 |
|    | RNA Hisat2-Stringtie analysis process        | 二代基因组测序即Next Generation Sequencing (NGS)是一种基于边合成边测序的方式。NGS在保持了测序高准确度的同时，大幅的提高了测序速度。目前NGS已经普遍的应用于全基因组测序、全外显子测序、转录组测序和表观遗传学等领域。<br>该流程以NGS得到的fast.gz文件作为输入，流程去接头、比对、转录组表达定量等步骤。                                                                                 |
|    | Human NGS data CNV detection                 | 部分癌症患者会出现基因的大片段增删突变，甚至染色体水平上的缺失或扩增性突变，因此Copy Number Variation(CNV)可以作为检测是否患癌的标志。本流程使用患者的NGS测序数据，进行CNV变异筛查，进而判断取样者是否有患癌风险。<br>该流程以CNVkit为核心，基于输入的fastq，以hg38人基因组数据生成的参考基因拷贝数分布为参考基线，能够自动的完成输入数据的比对排序，以及拷贝数分布计算，并输出可视化图表以供查看。                                   |

## 我的发布

在“资产市场>我的发布”中可以查看该用户所有发布资产的情况。对于发布成功的资产，可以在“组织共享”中查看和订阅。

- 对于发布成功的资产，可以进行下线操作，下线后已经订阅的资产将不可使用。

图 2-214 下线资产

| 名称            | 类型 | 版本  | 状态 | 创建时间                | 更新时间                | 状态  | 操作           |
|---------------|----|-----|----|---------------------|---------------------|-----|--------------|
| gppp01        | 流程 | 1.0 | -- | 2022/10/10 11:10:34 | 2022/10/10 11:11:18 | 已发布 | 编辑   删除   下线 |
| import-apt123 | 应用 | 1.0 | -- | 2022/10/09 18:16:10 | 2022/10/09 18:17:05 | 已发布 | 编辑   删除   下线 |

- 对于发布失败或者下线的资产，可以进行编辑，删除，发布等操作。

### 说明

- 在“我的发布”中，只能下线发布账号发布的资产。管理员可以下线“组织共享”中发布的所有资产。
- 发布资产前，需要先设置商标。详细可参考[商标设置](#)。

## 2.14.2 使用 Variant Calling Based On NGS 流程

二代基因组测序即Next Generation Sequencing (NGS)是一种基于边合成边测序的方式。NGS在保持了测序高准确度的同时，大幅的提高了测序速度。目前NGS已经普遍的应用于全基因组测序、全外显子测序、转录组测序表观遗传学等领域。

该流程以NGS得到的fastq作为输入，通过质控，比对，得到比对后的bam文件，及对fastq和bam文件的质控报告。

NGS流程由fastp、bwa-mem、picard-insertsize、bamqc应用构成，各应用说明如表2-24所示。

表 2-24 应用说明

| 应用名称              | 说明                                      |
|-------------------|-----------------------------------------|
| fastp             | 对测序得到的fastq数据进行质控。                      |
| bwa-mem           | 将质控之后得到的Clean Reads比对到参考基因组上。           |
| picard-insertsize | 针对构建Index后的bam文件，统计测序数据的Insert size的分布。 |
| bamqc             | 评估比对得到的bam文件的质量。                        |

使用NGS的详细步骤如下所示：

- [步骤1：订阅流程](#)
- [步骤2：上传待分析数据](#)
- [步骤3：创建分析作业](#)
- [步骤4：查看执行结果](#)

### 步骤 1：订阅流程

使用NGS流程前，需要您在资产市场中订阅该流程。

1. 在“资产市场”中查找“Variant Calling Based On NGS”流程。
2. 单击界面右侧“订阅”图标，订阅该流程。

订阅的流程将显示在“项目管理 > 工具”页面的流程列表中。

### 步骤 2：上传待分析数据

您在使用NGS流程前，请先上传待分析的数据，上传数据方法请参见[数据管理](#)。数据要求是成对的双端测序样本，即您需要上传两个样本数据文件，例如：xxx.R1.fastq.gz和xxx.R2.fastq.gz。

### 步骤 3：创建分析作业

1. 单击项目名称，进入“项目管理”页面，并选择“工具”页签。

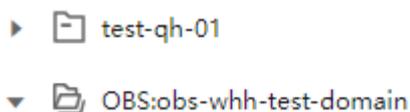
2. 单击ngs“操作”列“启动作业”，在弹出的新建作业页面，填写作业信息。“作业名称”、“标签”、和“描述”，设置“输出路径”、“优先级”、“计算节点标签”、“超时时间”和“加速类型”。
  - 基本信息：包含作业名称、标签、描述。
  - 输出路径：存放输出结果的路径，格式以/开头。例如项目中的output文件夹，输出路径可设置为/output。  
不填写路径时，默认以“作业名-UUID”格式生成输出路径。
  - 优先级：运行优先级，分为0~9级，优先级高的作业会被优先执行。默认值为0。

### 📖 说明

如果当前投递的IO加速的作业，超过sfs总的作业配额数，那么超出部分的作业会按照投递时间顺序，从最新投递的向前执行，不会按照优先级进行运行。

- 计算节点标签：作业会调度到含有相应标签的计算节点上。  
当应用也配置了标签，如果应用和作业的计算节点标签在同一计算节点上，则应用调度至该计算节点上；应用和它的作业，不管节点标签是否一致，都会被调度到应用的节点标签所对应的计算节点上。  
如果设置了不存在的计算节点标签，作业会进入等待，直至配置了相应的标签。
- 超时时间：作业运行时间超过设置时间时，认为超时，默认1440分钟，最大可设置为144000分钟，即作业运行至多100天。
- 加速类型：
  - 加速效率：IO加速>本地盘加速>无
  - 无：作业运行于OBS中，不使用加速。
  - IO加速：IO加速使用弹性文件服务（SFS）提供高性能的数据读写，作业运行时，会将非最终结果的数据存储在SFS中用以提高任务运行效率，作业执行完成后会清理释放SFS空间。对于涉及频繁读写场景的任务建议开启IO加速，开启前需要先[购买性能加速](#)。
  - 本地盘加速：使用计算节点的本地盘进行加速。使用本地盘加速时，需保证购买的[计算节点](#)带有“数据盘”。OBS桶中的数据不支持本地盘加速，使用OBS桶中数据用于本地盘加速，可能会导致作业运行失败。

图 2-215 “OBS” 标签代表数据引用来源为 OBS 桶



3. 作业信息填写完成后，单击“确定”，进入流程设计器页面。分别单击输入参数



图标，设置输入数据。

表 2-25 参数说明

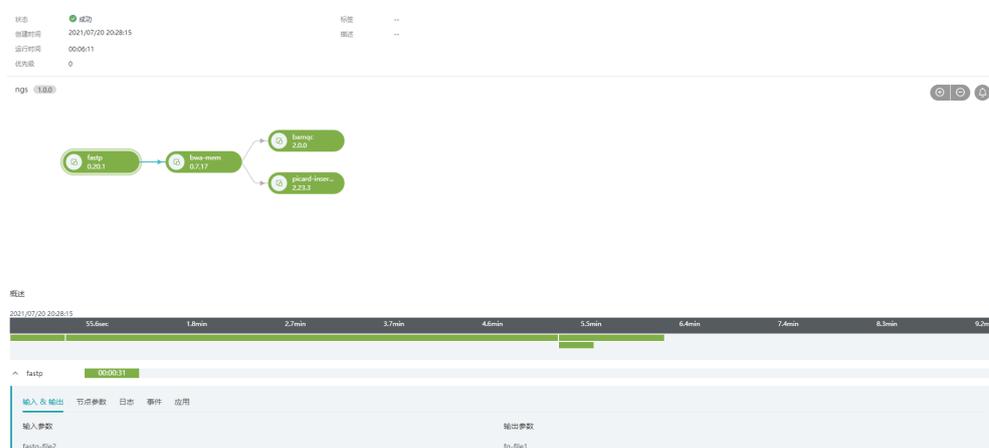
| 应用名称              | 参数   | 名称             | 类型        | 说明                      |
|-------------------|------|----------------|-----------|-------------------------|
| fastp             | 输入参数 | fastq-file1    | file      | 二代测序fastq的Read1文件。      |
|                   |      | fastq-file2    | file      | 二代测序fastq的Read2文件。      |
|                   | 输出参数 | fq-file1       | file      | Read1过滤之后输出fq.gz文件。     |
|                   |      | fq-file2       | file      | Read2过滤之后输出fq.gz文件      |
|                   |      | json-file      | file      | 以JSON文件的格式输出的质控报告。      |
|                   |      | html-file      | file      | 以HTML的格式输出易于阅读的质控报告。    |
| bwa-mem           | 输入参数 | fq-file1       | file      | 测序得到的fastq1文件。          |
|                   |      | fq-file2       | file      | 测序得到的fastq2文件。          |
|                   |      | ref-dir        | directory | 参考基因组序列。                |
|                   |      | seq-platform   | string    | 测序平台，如MGI、Illumina。     |
|                   |      | sample-id      | string    | 文件前缀，如NA12878。          |
|                   | 输出参数 | sorted-bam     | file      | 比对和排序之后得到的bam文件。        |
|                   |      | flagstat-file  | file      | 基于bam做统计。               |
| qualimap-bamqc    | 输入参数 | bam-file       | file      | 输入已经排序好的bam文件。          |
|                   | 输出参数 | out-dir        | directory | 质控报告的输出目录。              |
| picard-insertsize | 输入参数 | bam-file       | file      | 经过比对和排序之后得到的bam文件。      |
|                   |      | ref-dir        | directory | 参考基因组序列。                |
|                   | 输出参数 | insertsize-txt | file      | 输出的insert size分布的文本文件。  |
|                   |      | insertsize-pdf | file      | 输出的insert size分布的pdf文件。 |

4. 参数填写无误后，单击界面上方“启动作业”，运行作业。

## 步骤 4：查看执行结果

1. 在“项目管理”页面选择“作业”页签，单击创建的NGS作业名称，进入详情页，查看执行结果。

图 2-216 执行结果



2. 作业运行完成后，单击应用进度条中的“输出参数”，可跳转至数据所在路径，获取运行结果数据。

### 2.14.3 使用 Docking Summary 流程

分子对接（molecular docking）是分子模拟的重要方法之一，其本质是两个或多个分子之间的识别过程，其过程涉及分子之间的空间匹配和能量匹配。分子对接方法在药物设计、材料设计等领域有广泛的应用。

在药物分子产生药效反应的过程中，药物分子与靶酶相互结合，首先就需要两个分子充分接近，采取合适的取向，使两者在必要的部分相互契合，发生相互作用，继而通过适当的构象调整，得到一个稳定的复合物构象，通过分子对接确定复合物中的两个分子正确的相对位置和取向，研究两个分子的构象特别是底物构象在形成复合物过程中的变化，是确定药物作用机制，设计新药的基础。分子对接计算是把配体分子放在受体活性位点的位置，然后按照几何互补、能量互补以及化学环境互补的原则来实时评价配体与受体相互作用的好坏，并找到两个分子之间最佳的结合模式。由于分子对接考虑了受体结构的信息以及受体和药物分子之间的相互作用信息，因此从原理上讲，它比仅仅从配体结构出发的药物设计方法更加具有合理性。

Docking Summary流程由ligand-smiles-to-3dsdf、ligand-3dsdf-to-pdbqt、receptor-pdb-to-pdbqt、qvina-w、docking-summary应用构成，各应用说明如表2-26所示。

表 2-26 应用说明

| 应用名称                   | 说明                                   |
|------------------------|--------------------------------------|
| ligand-smiles-to-3dsdf | 将配体的smiles结构式转换为sdf文件。               |
| ligand-3dsdf-to-pdbqt  | 将配体的sdf文件转换为pdbqt文件。                 |
| receptor-pdb-to-pdbqt  | 将pdb文件转换为pdbqt文件。                    |
| qvina-w                | 执行分子对接操作（此处为blind docking，无需指定结合位点）。 |

| 应用名称            | 说明        |
|-----------------|-----------|
| docking-summary | 汇总分子对接结果。 |

使用Docking Summary的详细步骤如下所示：

- [步骤1：订阅流程](#)
- [步骤2：上传待分析数据](#)
- [步骤3：创建分析作业](#)
- [步骤4：查看执行结果](#)

## 步骤 1：订阅流程

使用Docking Summary流程前，需要您在资产市场中订阅该流程。

1. 在“资产市场”中查找“Docking Summary”流程。
2. 单击界面右侧“订阅”图标，订阅该流程。

订阅的流程将显示在“项目管理 > 工具”页面的流程列表中。

## 步骤 2：上传待分析数据

您在使用Docking Summary流程前，请先上传待分析的数据，上传数据方法请参见[数据管理](#)。

## 步骤 3：创建分析作业

1. 单击项目名称，进入“项目管理”页面，并选择“工具”页签。
2. 单击Docking Summary“操作”列“启动作业”，在弹出的新建作业页面，填写作业信息。“作业名称”、“标签”、和“描述”，设置“输出路径”、“优先级”、“计算节点标签”、“超时时间”和“加速类型”。
  - 基本信息：包含作业名称、标签、描述。
  - 输出路径：存放输出结果的路径，格式以/开头。例如项目中的output文件夹，输出路径可设置为/output。  
不填写路径时，默认以“作业名-UUID”格式生成输出路径。
  - 优先级：运行优先级，分为0~9级，优先级高的作业会被优先执行。默认值为0。

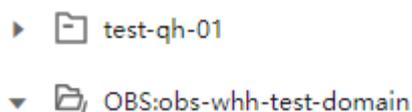
### 说明

如果当前投递的IO加速的作业，超过sfs总的作业配额数，那么超出部分的作业会按照投递时间顺序，从最新投递的向前执行，不会按照优先级进行运行。

- 计算节点标签：作业会调度到含有相应标签的计算节点上。  
当应用也配置了标签，如果应用和作业的计算节点标签在同一计算节点上，则应用调度至该计算节点上；应用和它的作业，不管节点标签是否一致，都会被调度到应用的节点标签所对应的计算节点上。  
如果设置了不存在的计算节点标签，作业会进入等待，直至配置了相应的标签。

- 超时时间：作业运行时间超过设置时间时，认为超时，默认1440分钟，最大可设置为144000分钟，即作业运行至多100天。
- 加速类型：  
加速效率：IO加速>本地盘加速>无
  - 无：作业运行于OBS中，不使用加速。
  - IO加速：IO加速使用弹性文件服务（SFS）提供高性能的数据读写，作业运行时，会将非最终结果的数据存储在SFS中用以提高任务运行效率，作业执行完成后会清理释放SFS空间。对于涉及频繁读写场景的任务建议开启IO加速，开启前需要先[购买性能加速](#)。
  - 本地盘加速：使用计算节点的本地盘进行加速。使用本地盘加速时，需保证购买的[计算节点](#)带有“数据盘”。OBS桶中的数据不支持本地盘加速，使用OBS桶中数据用于本地盘加速，可能会导致作业运行失败。

图 2-217 “OBS” 标签代表数据引用来源为 OBS 桶



3. 作业信息填写完成后，单击“确定”，进入流程设计器页面。分别单击输入参数

 图标，设置输入数据。

表 2-27 参数说明

| 应用名称                   | 参数   | 名称      | 类型        | 说明                                          |
|------------------------|------|---------|-----------|---------------------------------------------|
| ligand-smiles-to-3dsdf | 输入参数 | file-in | file      | 存放配体SMILES列表的文件，第一列为SMILES，第二列为配体名称，以Tab分隔。 |
|                        | 输出参数 | dir-out | directory | 转换后存放配体3D sdf文件的文件，默认文件名为3d.sdf。            |
| ligand-3dsdf-to-pdbqt  | 输入参数 | dir-in  | directory | 存放配体3D sdf文件的文件，默认文件名为3d.sdf。               |
|                        | 输出参数 | dir-out | directory | 转换后存放配体pdbqt文件的文件夹。                         |
| receptor-pdb-to-pdbqt  | 输入参数 | dir-in  | directory | 受体的pdb文件所在的文件夹。                             |
|                        | 输出参数 | dir-out | directory | 转换后的受体pdbqt文件所在的文件夹。                        |
| qvina-w                | 输入参数 | dir-rec | directory | 受体pdbqt文件所在文件夹。                             |
|                        |      | dir-lig | directory | 配体pdbqt文件所在文件夹。                             |
|                        | 输出参数 | dir-out | directory | 存放分析结果文件夹。                                  |

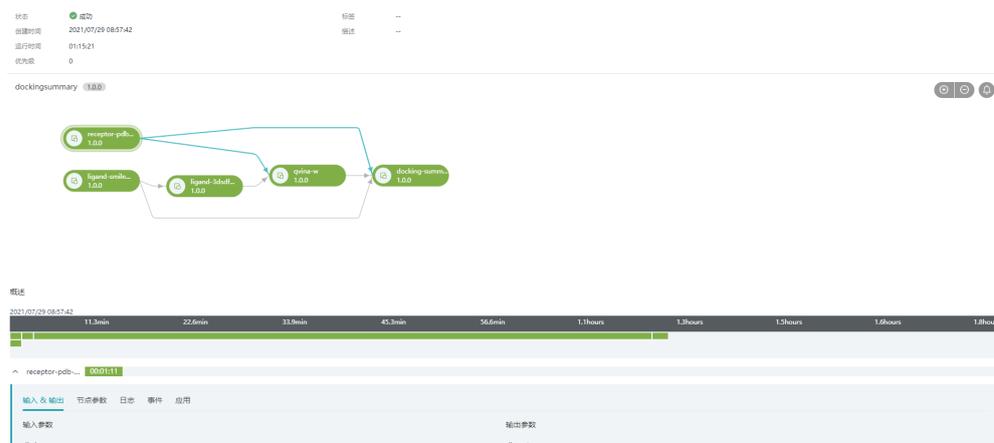
| 应用名称            | 参数   | 名称       | 类型        | 说明                                                    |
|-----------------|------|----------|-----------|-------------------------------------------------------|
| docking-summary | 输入参数 | dir-sdf  | directory | 配体SDF文件所在的文件夹。                                        |
|                 |      | dir-rec  | directory | 受体pdbqt文件所在的文件夹。                                      |
|                 |      | dir-dock | directory | 分子对接结果pdbqt文件所在的文件夹。                                  |
|                 | 输出参数 | dir-out  | directory | 分子对接结果汇总信息（矩阵.xlsx、3D可视化文件.pdb、配体及受体结构图片.svg等）所在的文件夹。 |

- 参数填写无误后，单击界面上方“启动作业”，运行作业。

## 步骤 4: 查看执行结果

- 在“项目管理”页面选择“作业”页签，单击创建的Docking Summary作业名称，进入详情页，查看执行结果。

图 2-218 执行结果



- 作业运行完成后，单击应用进度条中的“输出参数”，可跳转至数据所在路径，获取运行结果数据。

## 2.14.4 使用 RNA-Seq Analysis Based on STAR 流程

二代基因组测序即Next Generation Sequencing (NGS)是一种基于边合成边测序的方式。NGS在保持了测序高准确度的同时，大幅的提高了测序速度。目前NGS已经普遍的应用于全基因组测序、全外显子测序、转录组测序表观遗传学等领域。

该流程以NGS得到的fastq作为输入，通过质控，比对等步骤，输出针对fastq的qc报告，输出STAR比对得到的bam文件。

RNA-Seq流程由fastqc、trimmomatic、star应用构成，各应用说明如表2-28所示。

表 2-28 应用说明

| 应用名称        | 说明                                               |
|-------------|--------------------------------------------------|
| fastqc      | 对测序得到的fastq数据进行质控。                               |
| trimmomatic | 使用trimmomatic工具去掉测序接头，得到去接头后fastq格式的reads。       |
| star        | 使用STAR比对工具将去接头后的reads比对到参考基因组，得到bam文件和多种格式的统计报告。 |

使用RNA-Seq流程的详细步骤如下所示：

- [步骤1：订阅流程](#)
- [步骤2：上传待分析数据](#)
- [步骤3：创建分析作业](#)
- [步骤4：查看执行结果](#)

## 步骤 1：订阅流程

使用RNA-Seq流程前，需要您在资产市场中订阅该流程。

1. 在“资产市场”中查找“RNA-Seq Analysis Based on STAR”流程。
2. 单击界面右侧“订阅”图标，订阅该流程。  
订阅的流程将显示在“项目管理 > 工具”页面的流程列表中。

## 步骤 2：上传待分析数据

您在使用RNA-Seq流程前，请先上传待分析的数据，上传数据方法请参见[数据管理](#)。

## 步骤 3：创建分析作业

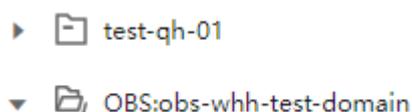
1. 单击项目名称，进入“项目管理”页面，并选择“工具”页签。
2. 在“RNA-Seq Analysis Based on STAR”流程操作列，单击“启动作业”，在弹出的新建作业页面，填写作业信息。“作业名称”、“标签”、和“描述”，设置“输出路径”、“优先级”、“计算节点标签”、“超时时间”和“加速类型”。
  - 基本信息：包含作业名称、标签、描述。
  - 输出路径：存放输出结果的路径，格式以/开头。例如项目中的output文件夹，输出路径可设置为/output。  
不填写路径时，默认以“作业名-UUID”格式生成输出路径。
  - 优先级：运行优先级，分为0~9级，优先级高的作业会被优先执行。默认值为0。

### 说明

如果当前投递的IO加速的作业，超过sfs总的作业配额数，那么超出部分的作业会按照投递时间顺序，从最新投递的向前执行，不会按照优先级进行运行。

- 计算节点标签：作业会调度到含有相应标签的计算节点上。  
当应用也配置了标签，如果应用和作业的计算节点标签在同一计算节点上，则应用调度至该计算节点上；应用和它的作业，不管节点标签是否一致，都会被调度到应用的节点标签所对应的计算节点上。  
如果设置了不存在的计算节点标签，作业会进入等待，直至配置了相应的标签。
- 超时时间：作业运行时间超过设置时间时，认为超时，默认1440分钟，最大可设置为144000分钟，即作业运行至多100天。
- 加速类型：  
加速效率：IO加速>本地盘加速>无
  - 无：作业运行于OBS中，不使用加速。
  - IO加速：IO加速使用弹性文件服务（SFS）提供高性能的数据读写，作业运行时，会将非最终结果的数据存储在SFS中用以提高任务运行效率，作业执行完成后会清理释放SFS空间。对于涉及频繁读写场景的任务建议开启IO加速，开启前需要先[购买性能加速](#)。
  - 本地盘加速：使用计算节点的本地盘进行加速。使用本地盘加速时，需保证购买的[计算节点](#)带有“数据盘”。OBS桶中的数据不支持本地盘加速，使用OBS桶中数据用于本地盘加速，可能会导致作业运行失败。

图 2-219 “OBS” 标签代表数据引用来源为 OBS 桶



3. 作业信息填写完成后，单击“确定”，进入流程设计器页面。分别单击输入参数  图标，设置输入数据。

表 2-29 参数说明

| 应用名称        | 参数   | 名称           | 类型        | 说明                         |
|-------------|------|--------------|-----------|----------------------------|
| fastqc      | 输入参数 | in-dir       | directory | fastq目录，包含待分析的fastq格式的文件。  |
|             |      | fastq-extend | string    | fastq文件后缀<extend>，适配不同的后缀。 |
| trimmomatic | 输入参数 | out-dir      | directory | 测序比对数据的质量控制报告的输出目录。        |
|             |      | fastq-extend | string    | fastq文件后缀<extend>，适配不同的后缀。 |
|             |      | in-dir       | directory | fastq目录，包含待分析的fastq格式的文件。  |
|             |      | para-str     | string    | 对扩展参数的支持，可以不填。             |

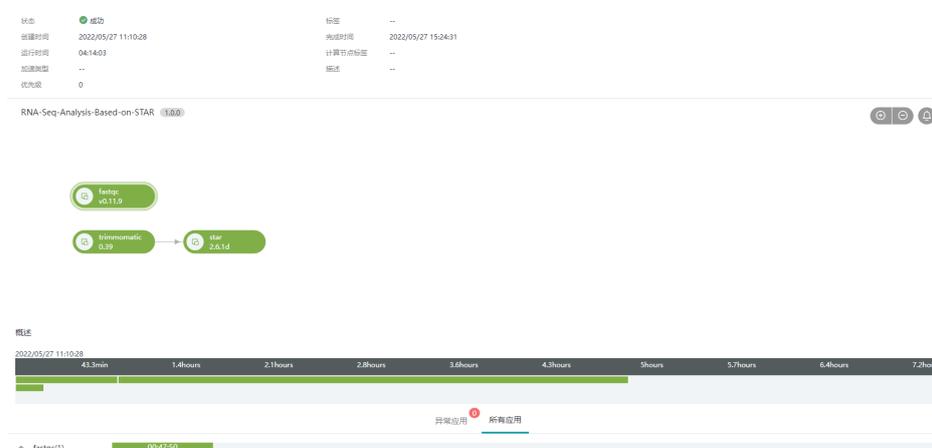
| 应用名称 | 参数          | 名称           | 类型             | 说明                                     |
|------|-------------|--------------|----------------|----------------------------------------|
|      |             | threads      | string         | 运行期线程数。                                |
|      |             | sample-name  | string         | sample name清单。                         |
|      | 输出参数        | out-dir      | directory      | trimmomatic去接头后的fastq输出目录。             |
| star | 输入参数        | in-dir       | directory      | fastq目录，包含待分析的fastq格式的文件。              |
|      |             | fastq-extend | string         | fastq文件后缀<extend>，适配不同的后缀。             |
|      |             | read-cmd     | string         | 读取输入文件时使用的命令，如读取.gz文件使用zcat，文本文件使用cat。 |
|      |             | para-str     | string         | 对扩展参数的支持，可以不填。                         |
|      |             | genome-dir   | directory      | 输入的参考基因组序列，已经使用STAR构建了index。           |
|      | sample-name | string       | sample name清单。 |                                        |
|      | 输出参数        | out-dir      | directory      | STAR比对结果所在文件夹。                         |

4. 参数填写无误后，单击界面上方“启动作业”，运行作业。

## 步骤 4：查看执行结果

1. 在“项目管理”页面选择“作业”页签，单击创建的RNA-Seq作业名称，进入详情页，查看执行结果。

图 2-220 执行结果



2. 作业运行完成后，单击应用进度条中的“输出参数”，可跳转至数据所在路径，获取运行结果数据。

## 2.14.5 使用 AutoGenome 镜像

AutoGenome是Notebook镜像，利用AutoML等技术帮助科研工作者在基因组学数据上端到端实现深度学习网络搜索，训练，评估，预测和解释的工具包。

使用AutoGenome镜像的详细步骤如下所示：

- [步骤1：订阅镜像](#)
- [步骤2：创建Notebook](#)
- [步骤3：预览AutoGenome案例](#)
- [步骤4：使用AutoGenome](#)

### 步骤 1：订阅镜像

使用AutoGenome镜像前，需要您在资产市场中订阅该镜像。

1. 登录医疗智能体，进入基因平台。
2. 在“资产市场”中查找“autogenome”镜像。
3. 单击界面右侧“订阅”图标，订阅该镜像。

订阅的镜像将显示在“项目管理 > 镜像”页面的镜像列表中。

### 步骤 2：创建 Notebook

1. 在“项目管理 > 开发”页面，单击“创建Notebook”，参考[表 参数说明](#)填写信息。

表 2-30 参数说明

| 参数名称 | 说明                                                                                                                                                                                                                                                            |
|------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 名称   | Notebook的名称。                                                                                                                                                                                                                                                  |
| 描述   | Notebook的简要描述。                                                                                                                                                                                                                                                |
| 镜像类型 | 选择“自定义”镜像。                                                                                                                                                                                                                                                    |
| 工作环境 | 选择“autogenome”镜像。                                                                                                                                                                                                                                             |
| CPU  | 设置CPU为8.0核。                                                                                                                                                                                                                                                   |
| GPU  | 设置GPU为1.0。                                                                                                                                                                                                                                                    |
| 内存   | 设置内存大于50G。                                                                                                                                                                                                                                                    |
| 存储路径 | <p>单击“存储路径”右侧  文件夹图标，设置用于存储 Notebook数据的OBS路径。如果想直接使用已有的文件或数据，可将数据提前上传至对应的OBS路径下。</p> <p>包含本项目桶最多挂载6个，不包含本项目桶最多挂载5个。</p> <p>用户在Notebook列表的所有文件读写操作是基于所选择的OBS路径下的内容操作。</p> |

2. 参数填写无误后，单击“立即创建”，创建Notebook。

### 步骤 3：预览 AutoGenome 案例

1. 打开创建的Notebook。
2. 在Notebook的根目录下的“AutoGenome-Examples”文件夹中，包含使用AutoGenome进行分析的示例，可供参考。

图 2-221 AutoGenome-Examples



表 2-31 AutoGenome 示例

| 示例名称                            | 说明                               |
|---------------------------------|----------------------------------|
| single_cell_rfcn_densenet.ipynb | 基于RFCN-DenseNet和表达谱对单细胞发育时期进行分类。 |
| pbmc_res_vae.ipynb              | 基于Res-VAE和表达谱对单细胞数据降维。           |

### 步骤 4：使用 AutoGenome

Notebook包含了端到端使用AutoGenome的代码，您可以使用Notebook案例复现AutoGenome示例的结果。

1. 以“pbmc\_res\_vae.ipynb”为例，用户可以打开相应的代码集，直接运行该Notebook，也可以调整代码集中的代码，进行二次开发。

图 2-222 基于 Res-VAE 和表达谱对单细胞数据降维

#### Autogenome实践 - 基于Res-VAE和表达谱对单细胞数据降维

##### 目录

1. 环境配置
2. 读取配置文件
3. 模型训练
4. 提取降维之后数据

##### 环境配置

获取该样例在Notebook容器内的上层文件夹名称

- 当使用OBS类型的notebook/JupyterLab实例时，当前目录默认为pbmc\_res\_vae

```
In [1]: ! import os
import sys
```

```
In [2]: ! try:
import tensorflow as tf
# 判断是否存在GPU, 此notebook需要用到gpu进行训练
if not tf.test.is_gpu_available():
    print("Error: The Notebook should run in a gpu environment.")
    sys.exit()
except:
    print("Error: TensorFlow-gpu can not be imported, the Notebook should run in a gpu environment.")
    sys.exit()
```

WARNING:tensorflow:From <ipython-input-2-4c76d339d1ee>:4: is\_gpu\_available (from tensorflow.python.framework.test\_util) is deprecated and will be removed in a future version.  
Instructions for updating:  
Use `tf.config.list\_physical\_devices('GPU')` instead.

2. 使用该Notebook时需要运行相应的代码模块，运行步骤如下所示。
  - a. 环境配置：加载AutoGenome以及辅助绘图的软件包。
  - b. 读取配置文件：通过json文件配置输入和输出路径。
  - c. 模型训练：针对提供的数据和模型参数，AutoGenome会搜索得到最优的神经网络结构。训练过程经过模型搜索阶段和模型训练阶段，在模型搜索阶段，根据json文件中的配置参数，对于选定的模型参数会训练一定步数，搜索得到较好结果的参数进行后续训练。训练过程中可选择在验证数据集上进行评估，评估结果更好的模型参数将会保留。
  - d. 提取降维之后数据：完成模型训练后，生成降维后的结果数据。

#### 📖 说明

当您在运行AutoGenome示例出现“Warning: restart the kernel and run the notebook again!”时，请单击Notebook工具栏中的按钮，重启Notebook环境，并重新执行出现告警的代码。

3. 您可以在Notebook工作目录中上传数据，使用AutoGenome工具。数据上传下载请参见[数据的上传和下载](#)。

#### 📖 说明

对于非挂载目录以外的目录下的文件，重启Notebook后会消失。例如，上传文件至Notebook的根目录下，该文件并不在被挂载的obs路径中，重启Notebook，该文件会消失。

图 2-223 Upload 上传数据



## 2.14.6 genomeagent 镜像

### 2.14.6.1 使用 genomeagent 镜像

智能体AI Agent，被设计为具有独立思考和行动能力的AI程序，只需提供一个研究目标 and 数据路径，即可自适应生成一个任务序列执行分析任务。

生信智能体AI Agent可以应对激增的多组学数据分析需求、高成本的生物信息学培训和满足不同应用场景的个性化需求，为多领域生命科学研究提供高效、定制化的数据分析支持。

使用genomeagent镜像的详细步骤如下所示：

- **步骤1：订阅镜像**
- **步骤2：创建Notebook**
- **步骤3：进入jupyterlab环境使用genomeagent智能体框架**

#### 步骤 1：订阅镜像

使用genomeagent镜像前，需要您在资产市场中订阅该镜像到项目中。

1. 登录医疗智能体，进入基因平台。
2. 在“资产市场”中查找“genomeagent”镜像。
3. 单击界面右侧“订阅”图标，订阅该镜像。  
订阅的镜像将显示在“项目管理 > 镜像”页面的镜像列表中。

## 步骤 2：创建 Notebook

1. 在“项目管理 > 项目名称 > 开发”页面，单击“创建Notebook”，参考[表 参数说明](#)填写信息。

表 2-32 参数说明

| 参数名称 | 说明                                                                                                                                                                                                                                |
|------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 名称   | Notebook的名称。                                                                                                                                                                                                                      |
| 描述   | Notebook的简要描述。                                                                                                                                                                                                                    |
| 镜像类型 | 选择“自定义”镜像。                                                                                                                                                                                                                        |
| 工作环境 | 选择“genomeagent”镜像。                                                                                                                                                                                                                |
| CPU  | 设置CPU为4.0核。                                                                                                                                                                                                                       |
| GPU  | 设置GPU为0。                                                                                                                                                                                                                          |
| 内存   | 设置内存大于8G。                                                                                                                                                                                                                         |
| 存储路径 | 单击“存储路径”右侧  文件夹图标，设置用于存储 Notebook数据的OBS路径。如果想直接使用已有的文件或数据，可将数据提前上传至对应的OBS路径下。<br>用户在Notebook列表的所有文件读写操作是基于所选择的当前项目OBS路径下的内容操作，暂不支持引用的项目数据路径。 |

2. 参数填写无误后，单击“立即创建”，创建Notebook。

## 步骤 3：进入 jupyterlab 环境使用 genomeagent 智能体框架

1. 选择[步骤2](#)中状态为“运行中”的Notebook实例，单击“操作”列的“打开”访问Notebook，请参考[打开Notebook](#)。
2. 进入Notebook的根目录下“genomeagent\_examples”文件夹中，包含使用“genomeagent\_examples”进行分析的示例，可供参考，以下均以example.ipynb教程示例进行。

表 2-33 genome-agent 示例

| 示例名称                        | 说明                                  |
|-----------------------------|-------------------------------------|
| example.ipynb               | jupyter形式的代码示例。                     |
| example_with_user_modify.py | 与example.ipynb内容基本相同，为python脚本形式示例。 |

- 配置eihealth-toolkit（见example.ipynb中配置提示，需要添加config以及切换到当前目录）。

```
[ ] 1 # 1. 添加health_config 添加命令提示初始配置。
2 | health_config add [Flags]
3 # 2. 使用health_config 添加命令提示初始配置。
4 | health_switch project [project_name]
```

- 初始化配置请参考**步骤3：初始化配置**章节。
- 切换项目，请参考**切换项目**，运行成功时会提示：  
health client config file url: /home/health-user/.health/config.ini  
add ak successfully!  
add sk successfully!  
add region successfully!  
add platform-id successfully!  
add user-name successfully!  
add password successfully!  
add domain-name successfully!  
add obs\_endpoint successfully!  
add obs\_install\_path successfully!  
switch to project xxxxx successfully!

### 📖 说明

镜像本身已默认安装health-linux-x86\_64版本，如需其他版本请参考eihealth-toolkit使用说明。

- 运行示例：

- 例如用户在/path/to/examples目录下存在以下数据文件和结构。镜像自带数据样例，位于~/work/genomeagent\_examples/case1。

```
$ls ~/work/genomeagent_examples
```

```
case1/
├── data
│   ├── mm39.fa
│   ├── mm39.ncbiRefSeq.gtf
│   ├── SRR1374921.fastq.gz
│   ├── SRR1374922.fastq.gz
│   ├── SRR1374923.fastq.gz
│   ├── SRR1374924.fastq.gz
│   └── TruSeq3-SE.fa
└── output
```

- 用户任务的设置配置为config\_data，见example.ipynb教程代码框。

```
# 用户输入任务相关参数以 key-value 字典形式提供
# data_list: 存储在平台项目的OBS中的数据文件路径，所在项目与当前开发环境项目需一致，建议为相对路径且在开发环境中当前python进程工作目录下能访问；以及提供自然语言表述的数据含义。
# output_dir: 存储在平台项目的OBS中的数据输出路径，所在项目与当前开发环境项目需一致，建议为相对路径且在开发环境中当前python进程工作目录下能访问；以及提供自然语言表述的数据含义。
# 任务目标goal_description，提供自然语言表述的生信任务目的
# 相关信息meta，可以提供额外信息给模型辅助完成任务，相关参数：
GenomeAgent(use_meta=True)即可纳入meta信息给模型，默认为False，不采用meta信息。
config_data = {
    # 必须，描述数据路径和数据含义
    'data_list': {
        "/case1/data/SRR1374921.fastq.gz": "single-end mouse rna-seq reads, replicate 1 in LoGlu group",
        "/case1/data/SRR1374922.fastq.gz": "single-end mouse rna-seq reads, replicate 2 in LoGlu group",
        "/case1/data/SRR1374923.fastq.gz": "single-end mouse rna-seq reads, replicate 1 in HiGlu group",
        "/case1/data/SRR1374924.fastq.gz": "single-end mouse rna-seq reads, replicate 2 in HiGlu group",
        "/case1/data/TruSeq3-SE.fa": "trimming adapter",
        "/case1/data/mm39.fa": "mouse mm39 genome fasta",
        "/case1/data/mm39.ncbiRefSeq.gtf": "mouse mm39 genome annotation"},
```

```
# 必需，指定该次任务的输出目录
'output_dir': './case1/output',
# 必需，描述该次任务的目的，例如find the differentially expressed genes
'goal_description': 'find the differentially expressed genes',
# 可选，提供给大模型的辅助参考meta信息，也可以完全为空，让大模型自主发挥
'meta': '\
step1, fastqc to check reads; \
step2, cutadapt with pair-end mode; \
step3, hisat2 build genome index and then alignment reads to mouse genome; \
step4, samtools view and sort for converting SAM to sorted BAM, and then samtools index
sorted BAM files; \
step5, htseq to get gene expression for each biological sample (the output of htseq: the
output of htseq: column 1 is gene id, column 2 is gene expression value, no header line); \
step6, calculate the differentially expressed genes using DESeq2 (conda install
bioconda::bioconductor-deseq2 -y)'
}
```

- c. 自动化任务设置，配置基因平台的作业信息、大模型访问信息，见 example.ipynb 教程代码框。

```
Agent_data = {
# 【必须】
# GenomeAgent相关参数，
# job_name可自定义
# project_name为当前项目名
# 【可选】
# 其余参数说明见：https://support.huaweicloud.com/clilist-eihealth/eihealth\_31\_0032.html
'job': {
# 【必须】
'job_name': 'case1',
'project_name': 'project_name',
# 【可选】
'nodeLabels': '',
'priority': '',
'health_tool': ''
},
# 【必需】
# 配合GenomeAgent(model='fromconfig')使用，配置模型的模型推理参数，预留参数接口，给
需要调用的自部署模型用
# 请注意：您调用的第三方大模型API服务，其数据传输及处理可能涉及第三方服务器。本应用及
平台不对第三方服务的内容、准确性、完整性及安全性承担任何责任。
# 请勿在使用过程中输入敏感、个人隐私或机密信息。使用本服务即表示您同意自行承担因第三方
API使用可能产生的风险，包括但不限于数据泄露、服务中断或结果不准确。
# 对于因使用第三方服务而导致的任何直接或间接损失，我们不承担任何法律责任。
# 如您调用的第三方大模型API服务需要设置APIKEY或者headers，请参考https://
platform.openai.com/docs/quickstart/step-2-set-up-your-api-key
# genomeagent默认从环境变量中获取'USER_APIKEY/USER_HEADERS'变量内容作为访问。

# 如果所用的大语言模型使用OpenAI模式部署的模型推理服务
# 模型参数attributes和网络访问proxy均应该遵循你的模型服务设置
# 例如:
'model_params':{'type': 'openai', 'url': 'https://api.deepseek.com/v1',
'model': 'deepseek-chat',
# 'proxy': "http://xxx.xxx.xxx.xxx:port",
'attributes': {'max_tokens': 4096}},
# 如果所用的大语言模型使用request模式部署的模型推理服务
# 请求体headers、模型参数attributes、网络访问proxy均应该遵循你的模型服务设置
# 例如:
# 'model_params':{'
# 'type': 'request', 'url': 'https://api.deepseek.com/chat/completions',
# 'model': 'deepseek-chat',
# 'proxy': {'http': "http://xxx.xxx.xxx.xxx:port", "https": "http://xxx.xxx.xxx.xxx:port"},
# 'attributes': {'max_tokens': 4096}},

# 基因平台工作流的APP参数，使用说明：https://support.huaweicloud.com/clilist-eihealth/
eihealth\_31\_0020.html
# 【必须】
# 'image': 'project_name/genomeagent:1.0.1'，该参数需要根据订阅的genomeagent修改，为
'image': '源项目名/镜像名:版本号'
# 【可选】
# 根据情况修改resources下的参数（例如cpu、memory和gpu）
```

```
# 其余参数无须改动
'app': {
  'image': 'project_name/genomeagent:1.0.0',
  # 根据情况修改resources下的参数（例如cpu、memory和gpu）
  'resources': {'cpu_type': 'X86', 'cpu': '4C', 'memory': '8G', 'gpu_type': '', 'gpu': '0'},
  # 其余参数无须改动
  'name': 'ga-app', 'version': 'v1.0', 'summary': '', 'description': '', 'labels': [],
  # 【不可改】
  # 注意commands与inputs字段下参数不可改动
  'commands': [ 'python3 .shell.py --cpu 0 --times 360 --interval 5 --script ${input-sh} && $
{input-sh}.log ' ],
  'inputs': [
    {enum': [], 'name': 'input-sh', 'description': '', 'required': False, 'concurrent': '', 'type':
FILE', 'pattern': '', 'values': []},
    {enum': [], 'name': 'input-dir', 'description': '', 'required': False, 'concurrent': '', 'type':
DIRECTORY', 'pattern': '', 'values': []},
    {enum': [], 'name': 'result-path', 'description': '', 'required': False, 'concurrent': '', 'type':
DIRECTORY', 'pattern': '', 'values': []}
  ],
  'outputs': [], 'node_labels': []
},
# 基因平台工作流的workflow参数，使用说明：https://support.huaweicloud.com/clilist-eihealth/eihealth\_31\_0026.html
# 【必须】
# 'app_id': 'ga-app::v1.0'，与APP参数中的'name': 'ga-app'和'version': 'v1.0'相匹配，默认相匹配，
# 无需改动
# 【可选】
# 仅需要根据情况修改tasks-resources下的参数（例如cpu、memory和gpu），其余参数无须改
# 动，注意tasks-inputs下参数不可改动
'workflow': {
  'tasks': [{
    'app_id': 'ga-app::v1.0',
    # 根据情况修改tasks-resources下的参数（例如cpu、memory和gpu）
    'resources': {'cpu': '4C', 'cpu_type': 'X86', 'memory': '8G', 'gpu_type': '', 'gpu': '0'},
    # 其余参数无须改动
    'task_name': 'ga',
    'display_name': '',
    'output_dir': '',
    'inputs': [
      {'name': 'input-sh', 'source': 'MANUAL', 'values': []},
      {'name': 'input-dir', 'source': 'MANUAL', 'values': []},
      {'name': 'result-path', 'source': 'MANUAL', 'values': []}
    ]
  }],
  'name': 'ga-wf', 'version': 'v1.0', 'summary': '', 'description': '', 'labels': [], 'timeout': 10080,
  'output_dir': ''
}
}}
config_data.update(Agent_data)
```

### 📖 说明

请注意：您调用的第三方大模型API服务，其数据传输及处理可能涉及第三方服务器。本应用及平台不对第三方服务的内容、准确性、完整性及安全性承担任何责任。请勿在使用过程中输入敏感、个人隐私或机密信息。使用本服务即表示您同意自行承担因第三方API使用可能产生的风险，包括但不限于数据泄露、服务中断或结果不准确。对于因使用第三方服务而导致的任何直接或间接损失，我们不承担任何法律责任。

- d. 将镜像中的genomeagent\_examples/case1示例文件拷贝到本开发环境所在项目的OBS桶中，见example.ipynb教程代码框。

```
# 获取本开发环境项目的OBS桶目录，复制 'examples' 文件夹并切换到目标目录
```

```
import os
import shutil
project_name = config_data['job']['project_name']
shutil.copytree(os.path.expanduser("~/work/genomeagent_examples"), os.path.expanduser(f"~/work/{project_name}/genomeagent_examples"), dirs_exist_ok=True)
os.chdir(os.path.expanduser(f"~/work/{project_name}/genomeagent_examples"))
os.getcwd()
```

- e. 加载GenomeAgent，见example.ipynb教程代码框。

```
# 如果有需要, 可以将以上配置保存为 YAML 文件
# import yaml
# with open('./case1.1/config.yaml', 'w') as outfile:
#     yaml.dump(config_data, outfile, default_flow_style=False, sort_keys=False)
from genomeagent import GenomeAgent
# 用config中model_params指定模型和接口等参数: model='fromconfig'
# 输出并执行代码: execute=True
# 用户不参与修改: user_modify=False (交互式用户修改目前仅支持python解释器或者python脚本下运行)
# 用EiFlow投递作业: local_mode=False
ga = GenomeAgent(config=config_data, model='fromconfig', execute=True, user_modify=False, local_mode=False, use_meta=True)
# 也可以从保存的文件中读取config
# ga = GenomeAgent(config='./case1.1/config.yaml', model='fromconfig', execute=True, user_modify=False, local_mode=False, use_meta=True)
# 开始执行自动任务
ga.run()
```

f. 开始执行自动任务, 见example.ipynb教程代码框。

图 2-224 大模型输出计划步骤

```
ga.run()
[Job Start]
[Round 0]
[Data]
meta_info
You should follow the related information to finish your goal: step1: fastqc to check reads; step2: cutadapt with pair-end mode; step3: hisat2 build genome index and then alignment reads to mouse genome; step4: samtools view and sort for converting SAM to sorted BAM, and then samtools index sorted BAM files; step5: htseq to get gene expression for each biological sample (the output of htseq: the output of htseq: column 1 is gene id, column 2 is gene expression value, no header line); step6: calculate the differentially expressed genes using DESeq2 (conda install bioconda:bioconductor-deseq2 -y)
history
[You have the following information in a list with the format file path: file description. I provide those files to you, so you don't need to prepare the data.]: ['case1/data/SRR1374921.fastq.gz', 'single-end mouse rna-seq reads, replicate 1 in L6L6U group', 'case1/data/SRR1374922.fastq.gz', 'single-end mouse rna-seq reads, replicate 2 in L6L6U group', 'case1/data/SRR1374923.fastq.gz', 'single-end mouse rna-seq reads, replicate 1 in H6L6U group', 'case1/data/SRR1374924.fastq.gz', 'single-end mouse rna-seq reads, replicate 2 in H6L6U group', 'case1/data/m39.fasta', 'mouse m39 genome fasta', 'case1/data/m39.ncbiRefSeq.gtf', 'mouse m39 genome annotation']
goal
find the differentially expressed genes
[AI]
plan
step 1: use FastQC to check the quality of single-end mouse rna-seq reads for each replicate (case1/data/SRR1374921.fastq.gz, case1/data/SRR1374922.fastq.gz, case1/data/SRR1374923.fastq.gz, case1/data/SRR1374924.fastq.gz)
step 2: use cutadapt in single-end mode to trim adapter sequences from the reads using the adapter file (case1/data/TruSeq-SE.fa) for each replicate (case1/data/SRR1374921.fastq.gz, case1/data/SRR1374922.fastq.gz, case1/data/SRR1374923.fastq.gz, case1/data/SRR1374924.fastq.gz)
step 3: use hisat2 to build a genome index using the mouse m39 genome fasta file (case1/data/m39.fasta) and then align the trimmed reads to the mouse genome for each replicate (case1/data/SRR1374921.fastq.gz, case1/data/SRR1374922.fastq.gz, case1/data/SRR1374923.fastq.gz, case1/data/SRR1374924.fastq.gz)
step 4: use samtools view to convert the SAM files generated from hisat2 Alignment to BAM format, then use samtools sort to sort the BAM files, and finally use samtools index to index the sorted BAM files for each replicate
step 5: use htseq-count to quantify gene expression for each biological sample using the sorted BAM files and the mouse m39 genome annotation file (case1/data/m39.ncbiRefSeq.gtf). The output should be formatted with column n 1 as gene id and column 2 as gene expression value, with no header line
step 6: use DESeq2 to calculate the differentially expressed genes between the L6L6U and H6L6U groups. Ensure to install DESeq2 using conda (conda install bioconda:bioconductor-deseq2 -y) before running the analysis
```

图 2-225 写出生信代码

```
[Round 1]
[Data]
meta_info
You should follow the related information to finish your goal: step1: fastqc to check reads; step2: cutadapt with pair-end mode; step3: hisat2 build genome index and then alignment reads to mouse genome; step4: samtools view and sort for converting SAM to sorted BAM, and then samtools index sorted BAM files; step5: htseq to get gene expression for each biological sample (the output of htseq: the output of htseq: column 1 is gene id, column 2 is gene expression value, no header line); step6: calculate the differentially expressed genes using DESeq2 (conda install bioconda:bioconductor-deseq2 -y)
history
Firstly, you have input with the format 'file path: file description' in a list: ['case1/data/SRR1374921.fastq.gz', 'single-end mouse rna-seq reads, replicate 1 in L6L6U group', 'case1/data/SRR1374922.fastq.gz', 'single-end mouse rna-seq reads, replicate 2 in L6L6U group', 'case1/data/SRR1374923.fastq.gz', 'single-end mouse rna-seq reads, replicate 1 in H6L6U group', 'case1/data/SRR1374924.fastq.gz', 'single-end mouse rna-seq reads, replicate 2 in H6L6U group', 'case1/data/m39.fasta', 'mouse m39 genome fasta', 'case1/data/m39.ncbiRefSeq.gtf', 'mouse m39 genome annotation', 'case1/output/all_outputs_should_be_written_under_this_dir']
You wrote a detailed plan to finish your goal. Your global goal is find the differentially expressed genes.
Your plan is as follows:
[step 1: use FastQC to check the quality of single-end mouse rna-seq reads for each replicate (case1/data/SRR1374921.fastq.gz, case1/data/SRR1374922.fastq.gz, case1/data/SRR1374923.fastq.gz, case1/data/SRR1374924.fastq.gz)
step 2: use cutadapt in single-end mode to trim adapter sequences from the reads using the adapter file (case1/data/TruSeq-SE.fa) for each replicate (case1/data/SRR1374921.fastq.gz, case1/data/SRR1374922.fastq.gz, case1/data/SRR1374923.fastq.gz, case1/data/SRR1374924.fastq.gz)
step 3: use hisat2 to build a genome index using the mouse m39 genome fasta file (case1/data/m39.fasta) and then align the trimmed reads to the mouse genome for each replicate (case1/data/SRR1374921.fastq.gz, case1/data/SRR1374922.fastq.gz, case1/data/SRR1374923.fastq.gz, case1/data/SRR1374924.fastq.gz)
step 4: use samtools view to convert the SAM files generated from hisat2 Alignment to BAM format, then use samtools sort to sort the BAM files, and finally use samtools index to index the sorted BAM files for each replicate
step 5: use htseq-count to quantify gene expression for each biological sample using the sorted BAM files and the mouse m39 genome annotation file (case1/data/m39.ncbiRefSeq.gtf). The output should be formatted with column n 1 as gene id and column 2 as gene expression value, with no header line
step 6: use DESeq2 to calculate the differentially expressed genes between the L6L6U and H6L6U groups. Ensure to install DESeq2 using conda (conda install bioconda:bioconductor-deseq2 -y) before running the analysis].
current task
step 1: use FastQC to check the quality of single-end mouse rna-seq reads for each replicate (case1/data/SRR1374921.fastq.gz, case1/data/SRR1374922.fastq.gz, case1/data/SRR1374923.fastq.gz, case1/data/SRR1374924.fastq.gz)
[AI Thinking...]
[AI Execute Code Start]
fastqc case1/data/SRR1374921.fastq.gz case1/data/SRR1374922.fastq.gz case1/data/SRR1374923.fastq.gz case1/data/SRR1374924.fastq.gz -o case1/output
health create job -u Babel74-a22c-11ef-b228-fa163ecb4f55 -n case1-step1 -i ga.input-sh-test1111:genomeagent_examples/case1/output_173164162939584116.sh.execute.sh;ga.input-dir=test1111:genomeagent_examples;pga.result-pat-h-test1111:genomeagent_examples/case1/output/
[The job case1-step1 have been submitted with id: 956d8ba-a302-11ef-b228-fa163ecb4f55]
[Execute Code Success], True, No error message
[Round 2]
[Data]
current task
step 1: use FastQC to check the quality of single-end mouse rna-seq reads for each replicate (case1/data/SRR1374921.fastq.gz, case1/data/SRR1374922.fastq.gz, case1/data/SRR1374923.fastq.gz, case1/data/SRR1374924.fastq.gz)
[AI Thinking...]
[AI Execute Code Start]
conda install bioconda:bioconductor-deseq2 -y
health create job -u Babel74-a22c-11ef-b228-fa163ecb4f55 -n case1-step6 -i ga.input-sh-test1111:genomeagent_examples/case1/output_173164162939584116.sh.execute.sh;ga.input-dir=test1111:genomeagent_examples;pga.result-pat-h-test1111:genomeagent_examples/case1/output/
[The job case1-step6 have been submitted with id: c7c8fac-a312-11ef-b228-fa163ecb4f55]
[Execute Code Success], True, No error message
[Round 6]
[Data]
[Job Finished]
see all scripts in case1/output_1731641629395841/all_step.sh and case1/output_1731641629395841/all_step.execute.sh
[AI] True
```

图 2-226 自动执行完成

```
Then, you finished the task: step 5: use htseq-count to quantify gene expression for each biological sample using the sorted BAM files and the mouse m39 genome annotation file (case1/data/m39.ncbiRefSeq.gtf). The output should be formatted with column 1 as gene id and column 2 as gene expression value, with no header line with followed codes:
htseq-count -f bam case1/output/SRR1374921.sorted.bam case1/data/m39.ncbiRefSeq.gtf -o case1/output/SRR1374921_gene_expression.txt
htseq-count -f bam case1/output/SRR1374922.sorted.bam case1/data/m39.ncbiRefSeq.gtf -o case1/output/SRR1374922_gene_expression.txt
htseq-count -f bam case1/output/SRR1374923.sorted.bam case1/data/m39.ncbiRefSeq.gtf -o case1/output/SRR1374923_gene_expression.txt
htseq-count -f bam case1/output/SRR1374924.sorted.bam case1/data/m39.ncbiRefSeq.gtf -o case1/output/SRR1374924_gene_expression.txt
You previously generated codes:
'''script''' = library(DESeq2); countData <- list.files("case1/output", pattern="*_gene_expression.txt", full.names=TRUE); countData <- lapply(countData, function(file) { data <- read.table(file, as.is=TRUE, header=TRUE, sep="\t", as.is=TRUE); res <- results(res); write.table(res, file="case1/output/DESeq2_results.txt", sep="\t", quote=FALSE, col.names=NA); }); However, your code has errors and you should fix them: The script failed due to the absence of the DESeq2 package. The error message "Error in library(DESeq2) : there is no package called 'DESeq2'" indicates that the DESeq2 package is not installed in the R environment. To resolve this issue, you need to install the DESeq2 package before running the script. You can install it using the following command in R: "install.packages('bioconductor'); BiocManager::install('DESeq2')". This will ensure that the DESeq2 package is available for use in your R environment.
current task
step 6: use DESeq2 to calculate the differentially expressed genes between the L6L6U and H6L6U groups. Ensure to install DESeq2 using conda (conda install bioconda:bioconductor-deseq2 -y) before running the analysis
[AI Thinking...]
[AI Execute Code Start]
conda install bioconda:bioconductor-deseq2 -y
health create job -u Babel74-a22c-11ef-b228-fa163ecb4f55 -n case1-step6 -i ga.input-sh-test1111:genomeagent_examples/case1/output_173164162939584116.sh.execute.sh;ga.input-dir=test1111:genomeagent_examples;pga.result-pat-h-test1111:genomeagent_examples/case1/output/
[The job case1-step6 have been submitted with id: c7c8fac-a312-11ef-b228-fa163ecb4f55]
[Execute Code Success], True, No error message
[Round 6]
[Data]
[Job Finished]
see all scripts in case1/output_1731641629395841/all_step.sh and case1/output_1731641629395841/all_step.execute.sh
[AI] True
```

g. 可以在jupyter的文件系统中看到运行结果, 也可以在基因平台项目的数据页面查看文件结果:

图 2-227 jupyter 中查看运行结果

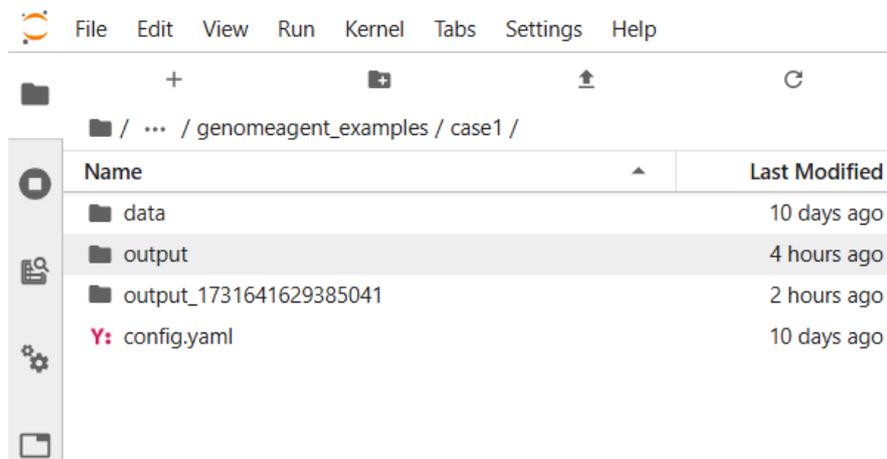
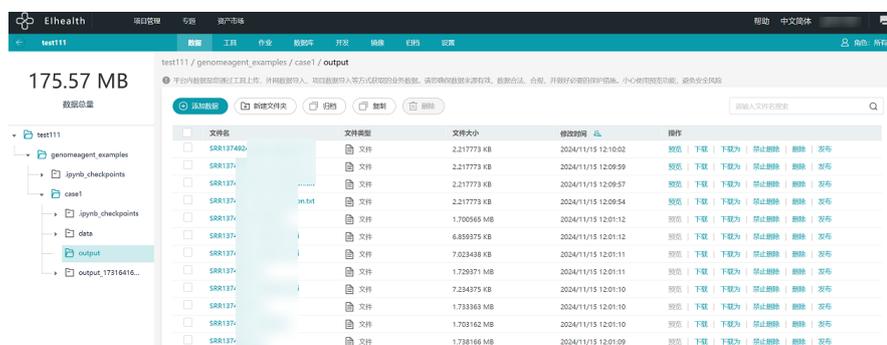


图 2-228 基因平台查看运行结果



h. 在基因平台项目中查看自动投递的作业。

图 2-229 查看自动投递的作业



## 2.14.6.2 genomeagent 镜像功能点介绍

此章节介绍genomeagent镜像的基本功能。

## 功能点 1: 自动生成任务计划、自动生成子步骤代码

```
[Job Start]
[Round 0]
[INFO]
[WARN]
[You have the following information in a list with the format file path: file description. I provide these files to you, so you don't need to prepare the data. ( 'examples/casel.1/data/SRR1374921.fastq.gz', 'single-end mouse rna-seq reads, replicate 1 in L601u group', 'examples/casel.1/data/SRR1374922.fastq.gz', 'single-end mouse rna-seq reads, replicate 2 in L601u group', 'examples/casel.1/data/SRR1374923.fastq.gz', 'single-end mouse rna-seq reads, replicate 1 in HG10 group', 'examples/casel.1/data/SRR1374924.fastq.gz', 'single-end mouse rna-seq reads, replicate 2 in HG10 group', 'examples/casel.1/data/TruSeq-SE.fa', 'trimming adapter', 'examples/casel.1/data/m29.fa', 'mouse m29 genome fasta', 'examples/casel.1/data/m29.ncbiRefSeq.gtf', 'mouse m29 genome annotation') ]
[You find the differentially expressed genes]
[All]
[Info]
[Step 1: Use Trimmomatic for quality control and adapter trimming on the single-end mouse rna-seq reads. This will involve specifying the adapter file (TruSeq-SE.fa) and the input fastq.gz files (SRR1374921.fastq.gz, SRR1374922.fastq.gz, SRR1374923.fastq.gz, SRR1374924.fastq.gz).]
[Step 2: Use HISAT2 for aligning the trimmed reads to the mouse m29 genome. This step will require the mouse genome fasta file (m29.fa) and the output from the Trimmomatic step.]
[Step 3: Use SAMtools to convert the HISAT2 output (SAM format) to BAM format, sort the BAM file, and index the sorted BAM file for efficient retrieval of alignments.]
[Step 4: Use featureCounts from the Subread package to count the reads that map to genes in the mouse m29 genome annotation (m29.ncbiRefSeq.gtf). This step will use the sorted BAM files from SAMtools as input.]
[Step 5: Use DESeq2 for differential expression analysis. This R package will analyze the count data from featureCounts, considering the replicates and groups (L601u vs. HG10) to identify differentially expressed genes.]
[Round 1]
[INFO]
[WARN]
[You have input with the format 'file path: file description' in a list: ( 'examples/casel.1/data/SRR1374921.fastq.gz', 'single-end mouse rna-seq reads, replicate 1 in L601u group', 'examples/casel.1/data/SRR1374922.fastq.gz', 'single-end mouse rna-seq reads, replicate 2 in L601u group', 'examples/casel.1/data/SRR1374923.fastq.gz', 'single-end mouse rna-seq reads, replicate 1 in HG10 group', 'examples/casel.1/data/SRR1374924.fastq.gz', 'single-end mouse rna-seq reads, replicate 2 in HG10 group', 'examples/casel.1/data/TruSeq-SE.fa', 'trimming adapter', 'examples/casel.1/data/m29.fa', 'mouse m29 genome fasta', 'examples/casel.1/data/m29.ncbiRefSeq.gtf', 'mouse m29 genome annotation', 'examples/casel.1/output/SRR1374921.fastq.gz', 'SRR1374921.fastq.gz', 'SRR1374922.fastq.gz', 'SRR1374923.fastq.gz', 'SRR1374924.fastq.gz').]
[You write a detailed plan to finish your goal. Your global goal is find the differentially expressed genes. This will involve specifying the adapter file (TruSeq-SE.fa) and the input fastq.gz files (SRR1374921.fastq.gz, SRR1374922.fastq.gz, SRR1374923.fastq.gz, SRR1374924.fastq.gz).]
[Step 1: Use Trimmomatic for quality control and adapter trimming on the single-end mouse rna-seq reads. This will involve specifying the adapter file (TruSeq-SE.fa) and the input fastq.gz files (SRR1374921.fastq.gz, SRR1374922.fastq.gz, SRR1374923.fastq.gz, SRR1374924.fastq.gz).]
[Step 2: Use HISAT2 for aligning the trimmed reads to the mouse m29 genome. This step will require the mouse genome fasta file (m29.fa) and the output from the Trimmomatic step.]
[Step 3: Use SAMtools to convert the HISAT2 output (SAM format) to BAM format, sort the BAM file, and index the sorted BAM file for efficient retrieval of alignments.]
[Step 4: Use featureCounts from the Subread package to count the reads that map to genes in the mouse m29 genome annotation (m29.ncbiRefSeq.gtf). This step will use the sorted BAM files from SAMtools as input.]
[Step 5: Use DESeq2 for differential expression analysis. This R package will analyze the count data from featureCounts, considering the replicates and groups (L601u vs. HG10) to identify differentially expressed genes.]
[Current task]
[Step 1: Use Trimmomatic for quality control and adapter trimming on the single-end mouse rna-seq reads. This will involve specifying the adapter file (TruSeq-SE.fa) and the input fastq.gz files (SRR1374921.fastq.gz, SRR1374922.fastq.gz, SRR1374923.fastq.gz, SRR1374924.fastq.gz).]
[AI Thinking...]
[AI Execute Code Start]
conda install -y trimmomatic

# Trimming for SRR1374921.fastq.gz
trimmomatic SE -phred33 examples/casel.1/data/SRR1374921.fastq.gz examples/casel.1/output/SRR1374921_trimmed.fastq.gz ILLUMINACLIP:examples/casel.1/data/TruSeq-SE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

# Trimming for SRR1374922.fastq.gz
trimmomatic SE -phred33 examples/casel.1/data/SRR1374922.fastq.gz examples/casel.1/output/SRR1374922_trimmed.fastq.gz ILLUMINACLIP:examples/casel.1/data/TruSeq-SE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

# Trimming for SRR1374923.fastq.gz
trimmomatic SE -phred33 examples/casel.1/data/SRR1374923.fastq.gz examples/casel.1/output/SRR1374923_trimmed.fastq.gz ILLUMINACLIP:examples/casel.1/data/TruSeq-SE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

# Trimming for SRR1374924.fastq.gz
trimmomatic SE -phred33 examples/casel.1/data/SRR1374924.fastq.gz examples/casel.1/output/SRR1374924_trimmed.fastq.gz ILLUMINACLIP:examples/casel.1/data/TruSeq-SE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

Do you want to modify codes? (y/n) (wait for 0.50 min)
skip modify codes

[The job case1-step1 have been submitted with id: ea977e4-596b-11ef-b222-fa163e6b4f55]
Check job status! /home/health/user/work/geno-research/genome-agent-dev-v0.3.0/health get job ea977e4-596b-11ef-b222-fa163e6b4f55 --detail
```

## 功能点 2: 自动提交 EFlow 代码任务

```
>>> from genomeagent import GenomeAgent as GenomeAgent
app not exists, now create app
create app success! app_id: 92325ed-596b-11ef-b222-fa163e6b4f55
workflow not exists, now create workflow
create workflow success! workflow_id: ea962922-596b-11ef-b222-fa163e6b4f55

[AI Thinking...]
[AI Execute Code Start]
conda install -y trimmomatic

# Trimming for SRR1374921.fastq.gz
trimmomatic SE -phred33 examples/casel.1/data/SRR1374921.fastq.gz examples/casel.1/output/SRR1374921_trimmed.fastq.gz ILLUMINACLIP:examples/casel.1/data/TruSeq-SE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

# Trimming for SRR1374922.fastq.gz
trimmomatic SE -phred33 examples/casel.1/data/SRR1374922.fastq.gz examples/casel.1/output/SRR1374922_trimmed.fastq.gz ILLUMINACLIP:examples/casel.1/data/TruSeq-SE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

# Trimming for SRR1374923.fastq.gz
trimmomatic SE -phred33 examples/casel.1/data/SRR1374923.fastq.gz examples/casel.1/output/SRR1374923_trimmed.fastq.gz ILLUMINACLIP:examples/casel.1/data/TruSeq-SE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

# Trimming for SRR1374924.fastq.gz
trimmomatic SE -phred33 examples/casel.1/data/SRR1374924.fastq.gz examples/casel.1/output/SRR1374924_trimmed.fastq.gz ILLUMINACLIP:examples/casel.1/data/TruSeq-SE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36

Do you want to modify codes? (y/n) (wait for 0.50 min)
skip modify codes

[The job case1-step1 have been submitted with id: ea977e4-596b-11ef-b222-fa163e6b4f55]
Check job status! /home/health/user/work/geno-research/genome-agent-dev-v0.3.0/health get job ea977e4-596b-11ef-b222-fa163e6b4f55 --detail
```

## 功能点 3: 报错时自动修正

### 报错时自动修正

```
[Current task]
[Step 9: Add or replace read groups in the BAM file using Picard]
[AI Thinking...]
[AI Execute Code Start]
picard AddOrReplaceReadGroups I=examples/caseWS/output/test.marked_duplicates.bam O=examples/caseWS/output/test_read_group.bam RGID=4 RGLB=11b1 RGPL=illumina RPFU=unit1 RGSZ=20

Do you want to modify codes? (y/n) (wait for 0.50 min)
skip modify codes

[The job case-ws-step9 have been submitted with id: c804da-596b-11ef-b222-fa163e6b4f55]
Check job status! /home/health/user/work/geno-research/genome-agent-dev-v0.3.0/health get job c804da-596b-11ef-b222-fa163e6b4f55 --detail

[Step failed with log:
[stderr] /home/geno-research/genome-agent-dev-v0.3.0-platform/agent_codes/genomeagent/examples/caseWS/output/9.sh.execute.sh: line 11: picard: command not found
[stderr] Bash codes failed with exit code: 127
[stderr] /home/geno-research/genome-agent-dev-v0.3.0-platform/agent_codes/genomeagent/examples/caseWS/output/9.sh.execute.sh: line 11: picard: command not found
[stderr] Bash codes failed with exit code: 127
[AI suggestion to fix]
The Bash script execution failed due to the 'picard: command not found' error, which indicates that the 'picard' command is not recognized in the environment where the script was executed. This could be because the Picard tools are not installed in the system, or they are not included in the system's PATH environment variable, preventing the shell from locating the executable. To resolve this issue, ensure that Picard tools are properly installed on the system and that their installation directory is added to the PATH environment variable. If Picard tools are installed but not accessible, you might need to specify the full path to the Picard executable in your script. Additionally, the exit code 127 typically signifies that the command was not found, confirming the diagnosis.
[Execute Code Failed 1/25 times at current step.]

[Current task]
[Step 9: Add or replace read groups in the BAM file using Picard]
[AI Thinking...]
[AI Execute Code Start]
conda install -y picard && picard AddOrReplaceReadGroups I=examples/caseWS/output/test.marked_duplicates.bam O=examples/caseWS/output/test_read_group.bam RGID=4 RGLB=11b1 RGPL=illumina RPFU=unit1 RGSZ=20

Do you want to modify codes? (y/n) (wait for 0.50 min)
skip modify codes

[The job case-ws-step9 have been submitted with id: 01da8db-596b-11ef-b222-fa163e6b4f55]
Check job status! /home/health/user/work/geno-research/genome-agent-dev-v0.3.0/health get job 01da8db-596b-11ef-b222-fa163e6b4f55 --detail
```

### 报错时用户介入修正。

```
[AI Thinking...]
[AI Execute Code Start]
conda install -y picard && picard AddOrReplaceReadGroups I=examples/caseWS/output/test.marked_duplicates.bam O=examples/caseWS/output/test_read_group.bam RGID=4 RGLB=11b1 RGPL=illumina RPFU=unit1 RGSZ=20

Do you want to modify codes? (y/n) (wait for 0.50 min)
skip modify codes

[The job case-ws-step9 have been submitted with id: 01da8db-596b-11ef-b222-fa163e6b4f55]
Check job status! /home/health/user/work/geno-research/genome-agent-dev-v0.3.0/health get job 01da8db-596b-11ef-b222-fa163e6b4f55 --detail
```

## 功能点 4: 特定步骤中断 (人为 or 非人为), 前序步骤状态继承, 重新执行中断步骤

```
[Round 3]
[USER]
history
Firstly, you have input with the format 'file path; file description' in a list: ['examples/case1/data/SRR1374921.fastq.gz: single-end mouse rna-seq reads, replicate 1 in LoLu group', 'examples/case1/data/SRR1374922.fastq.gz: single-end mouse rna-seq reads, replicate 2 in LoLu group', 'examples/case1/data/TruSeq3-SE.fa: trimming adapter', 'examples/case1/data/m39.fa: mouse m39 genome fasta', 'examples/case1/data/m39.ncbiRefSeq.gtf: mouse m39 genome annotation', 'examples/case1/output: all output should be stored under this dir'].
You wrote a detailed plan to finish your goal. Your global goal is find the differentially expressed genes.
Your plan is [Step 1: Use Trimmomatic for quality control and adapter trimming of the raw RNA-seq reads.
Step 2: Use FastQC to assess the quality of the trimmed reads.
Step 3: Use HISAT2 for aligning the trimmed reads to the mouse m39 genome.
Step 4: Use SAMtools to sort and index the aligned BAM files.
Step 5: Use featureCounts from the Subread package to count reads mapping to genes based on the m39 genome annotation.
Step 6: Use DESeq2 for differential gene expression analysis between the LoLu and HiLu groups.
Step 7: Use R to generate visualizations such as MA plots and heatmaps to interpret the differential expression results.].
Then, you finished the task: Step 1: Use Trimmomatic for quality control and adapter trimming of the raw RNA-seq reads, with followed codes:
conda install -y trimmomatic

# Trimming reads for SRR1374921
trimmomatic SE -phred35 examples/case1/data/SRR1374921.fastq.gz examples/case1/output/Tracback (most recent call last):
File "conda", line 1, in <module>
File "/home/health-user/miniconda/lib/python3.10/site-packages/genomeagent/agent.py", line 654, in run
run_code_generation_phase(status = self.run_code_generation_phase())
File "/home/health-user/miniconda/lib/python3.10/site-packages/genomeagent/agent.py", line 600, in run_code_generation_phase
INFO_STR_USER = self.generator.format_user_prompt(prompt, self.global_round, self.gul_path, skip["input"])
File "/home/health-user/miniconda/lib/python3.10/site-packages/genomeagent/prompt.py", line 190, in format_user_prompt
self.slow_print(prompt[key], speed=0.001)
File "/home/health-user/miniconda/lib/python3.10/site-packages/genomeagent/prompt.py", line 176, in slow_print
time.sleep(speed)
KeyboardInterrupt
>>> ag_1
ag_load_state( ag_local_model_engines
>>> ag_load_state(3)
>>> ag_run(0)
[Round 2]
[USER]
history
Firstly, you have input with the format 'file path; file description' in a list: ['examples/case1/data/SRR1374921.fastq.gz: single-end mouse rna-seq reads, replicate 1 in LoLu group', 'examples/case1/data/SRR1374922.fastq.gz: single-end mouse rna-seq reads, replicate 2 in LoLu group', 'examples/case1/data/TruSeq3-SE.fa: trimming adapter', 'examples/case1/data/m39.fa: mouse m39 genome fasta', 'examples/case1/data/m39.ncbiRefSeq.gtf: mouse m39 genome annotation', 'examples/case1/output: all output should be stored under this dir'].
You wrote a detailed plan to finish your goal. Your global goal is find the differentially expressed genes.
Your plan is [Step 1: Use Trimmomatic for quality control and adapter trimming of the raw RNA-seq reads.
Step 2: Use FastQC to assess the quality of the trimmed reads.
Step 3: Use HISAT2 for aligning the trimmed reads to the mouse m39 genome.
Step 4: Use SAMtools to sort and index the aligned BAM files.
Step 5: Use featureCounts from the Subread package to count reads mapping to genes based on the m39 genome annotation.
Step 6: Use DESeq2 for differential gene expression analysis between the LoLu and HiLu groups.
Step 7: Use R to generate visualizations such as MA plots and heatmaps to interpret the differential expression results.].
Then, you finished the task: Step 1: Use Trimmomatic for quality control and adapter trimming of the raw RNA-seq reads, with followed codes:
conda install -y trimmomatic

current task
Step 3: Use HISAT2 for aligning the trimmed reads to the mouse m39 genome.
[AI Thinking...]
[AI Execute Code Start]
conda install -y hisat2

# Building HISAT2 index for the mouse m39 genome
hisat2-build examples/case1/data/m39.fa examples/case1/output/m39_hisat2_index

# Aligning trimmed reads for SRR1374921 to the mouse m39 genome
hisat2 -x examples/case1/output/m39_hisat2_index -U examples/case1/output/SRR1374921_trimmed.fastq.gz -S examples/case1/output/SRR1374921_aligned.sam

# Aligning trimmed reads for SRR1374922 to the mouse m39 genome
hisat2 -x examples/case1/output/m39_hisat2_index -U examples/case1/output/SRR1374922_trimmed.fastq.gz -S examples/case1/output/SRR1374922_aligned.sam

# Aligning trimmed reads for SRR1374923 to the mouse m39 genome
hisat2 -x examples/case1/output/m39_hisat2_index -U examples/case1/output/SRR1374923_trimmed.fastq.gz -S examples/case1/output/SRR1374923_aligned.sam

# Aligning trimmed reads for SRR1374924 to the mouse m39 genome
hisat2 -x examples/case1/output/m39_hisat2_index -U examples/case1/output/SRR1374924_trimmed.fastq.gz -S examples/case1/output/SRR1374924_aligned.sam

Do you want to modify codes? (y/n) (wait for 0.50 mins)
skip modify codes
[Round 3 finish]
```

## 功能点 5: 手动中断当前状态, 进行修改或加载其他步骤状态

- 中断修改计划

```
[USER]
input
['You have the following information in a list with the format file path; file description. I provide these files to you, so you don't need to prepare the data. ['examples/case1/data/SRR1374921.fastq.gz: single-end mouse rna-seq reads, replicate 1 in LoLu group', 'examples/case1/data/SRR1374922.fastq.gz: single-end mouse rna-seq reads, replicate 2 in LoLu group', 'examples/case1/data/TruSeq3-SE.fa: trimming adapter', 'examples/case1/data/m39.fa: mouse m39 genome fasta', 'examples/case1/data/m39.ncbiRefSeq.gtf: mouse m39 genome annotation']
goal
find the differentially expressed genes

[AI]
plan
Step 1: Quality Control
Use FastQC to perform quality control checks on the raw sequencing data files: 'examples/case1/data/SRR1374921.fastq.gz', 'examples/case1/data/SRR1374922.fastq.gz', 'examples/case1/data/SRR1374923.fastq.gz', 'examples/case1/data/SRR1374924.fastq.gz'.
Step 2: Adapter Trimming
Use Trim Galore to trim the adapter sequences from the raw reads using the adapter file 'examples/case1/data/TruSeq3-SE.fa'.
Step 3: Quality Check
Re-run FastQC on the trimmed sequencing data to ensure the quality has improved and adapters have been successfully removed.
Step 4: Read Alignment
Use HISAT2 to align the trimmed reads to the mouse genome 'examples/case1/data/m39.fa'.
Step 5: Sort and Index BAM Files
Use SAMtools to sort and index the aligned BAM files generated from HISAT2 alignment.
Step 6: Quantification
Use featureCounts to quantify the gene expression levels based on the sorted BAM files and the genome annotation file 'examples/case1/data/m39.ncbiRefSeq.gtf'.
Step 7: Differential Expression Analysis
Use DESeq2 to perform differential expression analysis on the gene expression counts obtained from featureCounts, comparing the LoLu group (replicates 1 and 2) against the HiLu group (replicates 1 and 2).
Step 8: Result Visualization
Use R packages such as ggplot2 and pheatmap to visualize the results of the differential expression analysis, including MA plots, volcano plots, and heatmaps.

Do you want to modify plan? (y/n) (wait for 0.50 mins)
Please type the manually modified plan (format: single step in single line, multiple steps should have multiple lines, and with a separate line of '#done'):
Step 1: Quality Control. Use FastQC to perform quality control checks on the raw sequencing data files: 'examples/case1/data/SRR1374921.fastq.gz', 'examples/case1/data/SRR1374922.fastq.gz', 'examples/case1/data/SRR1374923.fastq.gz', 'examples/case1/data/SRR1374924.fastq.gz'.
Step 2: Adapter Trimming. Use cutadapt to trim the adapter sequences from the raw reads using the adapter file 'examples/case1/data/TruSeq3-SE.fa'.
Step 3: Quality Check. Re-run FastQC on the trimmed sequencing data to ensure the quality has improved and adapters have been successfully removed.
Step 4: Read Alignment. Use HISAT2 to align the trimmed reads to the mouse genome 'examples/case1/data/m39.fa'.
Step 5: Sort and Index BAM Files.
Use SAMtools to sort and index the aligned BAM files generated from HISAT2 alignment.
Step 6: Quantification. Use featureCounts to quantify the gene expression levels based on the sorted BAM files and the genome annotation file 'examples/case1/data/m39.ncbiRefSeq.gtf'.
Step 7: Differential Expression Analysis. Use DESeq2 to perform differential expression analysis on the gene expression counts obtained from featureCounts, comparing the LoLu group (replicates 1 and 2) against the HiLu group (replicates 1 and 2).
Step 8: Result Visualization. Use R packages such as ggplot2 and pheatmap to visualize the results of the differential expression analysis, including MA plots, volcano plots, and heatmaps.
#done

[Running user modified plan]
Step 1: Quality Control. Use FastQC to perform quality control checks on the raw sequencing data files: 'examples/case1/data/SRR1374921.fastq.gz', 'examples/case1/data/SRR1374922.fastq.gz', 'examples/case1/data/SRR1374923.fastq.gz', 'examples/case1/data/SRR1374924.fastq.gz'.
Step 2: Adapter Trimming. Use cutadapt to trim the adapter sequences from the raw reads using the adapter file 'examples/case1/data/TruSeq3-SE.fa'.
Step 3: Quality Check. Re-run FastQC on the trimmed sequencing data to ensure the quality has improved and adapters have been successfully removed.
Step 4: Read Alignment. Use HISAT2 to align the trimmed reads to the mouse genome 'examples/case1/data/m39.fa'.
Step 5: Sort and Index BAM Files.
Use SAMtools to sort and index the aligned BAM files generated from HISAT2 alignment.
Step 6: Quantification. Use featureCounts to quantify the gene expression levels based on the sorted BAM files and the genome annotation file 'examples/case1/data/m39.ncbiRefSeq.gtf'.
Step 7: Differential Expression Analysis. Use DESeq2 to perform differential expression analysis on the gene expression counts obtained from featureCounts, comparing the LoLu group (replicates 1 and 2) against the HiLu group (replicates 1 and 2).
Step 8: Result Visualization. Use R packages such as ggplot2 and pheatmap to visualize the results of the differential expression analysis, including MA plots, volcano plots, and heatmaps.
```

- 中断修改代码

```
[Round 4]
[USER]
history
Firstly, you have input with the format 'file path: file description' in a list: ['example/case1/data/SRR1374921.fastq.gz: single-end mouse rna-seq reads, replicate 1 in LoGlu group', 'example/case1/data/SRR1374922.fastq.gz: single-end mouse rna-seq reads, replicate 2 in LoGlu group', 'example/case1/data/SRR1374923.fastq.gz: single-end mouse rna-seq reads, replicate 1 in HiGlu group', 'example/case1/data/SRR1374924.fastq.gz: single-end mouse rna-seq reads, replicate 2 in HiGlu group', 'example/case1/data/TruSeq-SE.fa: trimming adapter', 'example/case1/data/m39.fa: mouse m39 genome fasta', 'example/case1/data/m39.ncbiRefSeq.gtf: mouse m39 genome annotation', 'example/case1/output: all outputs should be stored under this dir'].
You wrote a detailed plan to finish your goal. Your global goal is find the differentially expressed genes.
Your plan is [Step 1: Use TrinGalore for quality and adapter trimming of the raw RNA-seq reads.
Step 2: Use HISAT2 for aligning the trimmed reads to the mouse m39 genome.
Step 3: Use SAMtools to sort and index the aligned BAM files.
Step 4: Use featureCounts from the Subread package to count reads mapping to genes based on the m39.ncbiRefSeq.gtf annotation file.
Step 5: Use DESeq2 in R to perform differential expression analysis on the gene count data, identifying differentially expressed genes between the LoGlu and HiGlu groups.].
Then, you finished the task. Step 1: Use TrinGalore for quality and adapter trimming of the raw RNA-seq reads, with followed codes:
trin_galore --quality 20 --phred33 --stringency 1 --length 20 --output_dir example/case1/output/trimmed --adapter file:example/case1/data/TruSeq-SE.fa example/case1/data/SRR1374921.fastq.gz
trin_galore --quality 20 --phred33 --stringency 1 --length 20 --output_dir example/case1/output/trimmed --adapter file:example/case1/data/TruSeq-SE.fa example/case1/data/SRR1374922.fastq.gz
trin_galore --quality 20 --phred33 --stringency 1 --length 20 --output_dir example/case1/output/trimmed --adapter file:example/case1/data/TruSeq-SE.fa example/case1/data/SRR1374923.fastq.gz
trin_galore --quality 20 --phred33 --stringency 1 --length 20 --output_dir example/case1/output/trimmed --adapter file:example/case1/data/TruSeq-SE.fa example/case1/data/SRR1374924.fastq.gz
Then, you finished the task. Step 2: Use HISAT2 for aligning the trimmed reads to the mouse m39 genome, with followed codes:
hisat2 -x example/case1/data/m39.fa -i example/case1/output/trimmed/SRR1374921.fastq.gz -S example/case1/output/SRR1374921.sam
hisat2 -x example/case1/data/m39.fa -i example/case1/output/trimmed/SRR1374922.fastq.gz -S example/case1/output/SRR1374922.sam
hisat2 -x example/case1/data/m39.fa -i example/case1/output/trimmed/SRR1374923.fastq.gz -S example/case1/output/SRR1374923.sam
hisat2 -x example/case1/data/m39.fa -i example/case1/output/trimmed/SRR1374924.fastq.gz -S example/case1/output/SRR1374924.sam
Then, you finished the task. Step 3: Use SAMtools to sort and index the aligned BAM files, with followed codes:
samtools sort -o example/case1/output/SRR1374921.bam example/case1/output/SRR1374921.sam
samtools index example/case1/output/SRR1374921.bam
samtools sort -o example/case1/output/SRR1374922.bam example/case1/output/SRR1374922.sam
samtools index example/case1/output/SRR1374922.bam
samtools sort -o example/case1/output/SRR1374923.bam example/case1/output/SRR1374923.sam
samtools index example/case1/output/SRR1374923.bam
samtools sort -o example/case1/output/SRR1374924.bam example/case1/output/SRR1374924.sam
samtools index example/case1/output/SRR1374924.bam
current task
Step 4: Use featureCounts from the Subread package to count reads mapping to genes based on the m39.ncbiRefSeq.gtf annotation file.
[AI Thinking...]
[AI Execute Code Start]
featureCounts -a example/case1/data/m39.ncbiRefSeq.gtf -o example/case1/output/SRR1374921_featureCounts.txt example/case1/output/SRR1374921.bam
featureCounts -a example/case1/data/m39.ncbiRefSeq.gtf -o example/case1/output/SRR1374922_featureCounts.txt example/case1/output/SRR1374922.bam
featureCounts -a example/case1/data/m39.ncbiRefSeq.gtf -o example/case1/output/SRR1374923_featureCounts.txt example/case1/output/SRR1374923.bam
featureCounts -a example/case1/data/m39.ncbiRefSeq.gtf -o example/case1/output/SRR1374924_featureCounts.txt example/case1/output/SRR1374924.bam
Do you want to modify codes? (y/n) (wait for 0.50 mins)
skip modify codes
[Round 4 finish]
[Traceback (most recent call last):
File "conda", line 1, in <module>
File "/home/healthuser/miniconda/lib/python3.10/site-packages/genomewgmt/agent.py", line 655, in run
run_code_generation_phase_status = self.run_code_generation_phase()
File "/home/healthuser/miniconda/lib/python3.10/site-packages/genomewgmt/agent.py", line 644, in run_code_generation_phase
time.sleep(5)
KeyboardInterrupt
>>> sg.dbsql_prompt()
S
>>> sg.load_state(2)
>>> sg.run()
[Round 2]
[USER]
history
Firstly, you have input with the format 'file path: file description' in a list: ['example/case1/data/SRR1374921.fastq.gz: single-end mouse rna-seq reads, replicate 1 in LoGlu group', 'example/case1/data/SRR1374922.fastq.gz: single-end mouse rna-seq reads, replicate 2 in LoGlu group', 'example/case1/data/SRR1374923.fastq.gz: single-end mouse rna-seq reads, replicate 1 in HiGlu group', 'example/case1/data/SRR1374924.fastq.gz: single-end mouse rna-seq reads, replicate 2 in HiGlu group', 'example/case1/data/TruSeq-SE.fa: trimming adapter', 'example/case1/data/m39.fa: mouse m39 genome fasta', 'example/case1/data/m39.ncbiRefSeq.gtf: mouse m39 genome annotation', 'example/case1/output: all outputs should be stored under this dir'].
You wrote a detailed plan to finish your goal. Your global goal is find the differentially expressed genes.
Your plan is [Step 1: Use TrinGalore for quality and adapter trimming of the raw RNA-seq reads.
Step 2: Use HISAT2 for aligning the trimmed reads to the mouse m39 genome.
Step 3: Use SAMtools to sort and index the aligned BAM files.
Step 4: Use featureCounts from the Subread package to count reads mapping to genes based on the m39.ncbiRefSeq.gtf annotation file.
Step 5: Use DESeq2 in R to perform differential expression analysis on the gene count data, identifying differentially expressed genes between the LoGlu and HiGlu groups.].
Then, you finished the task. Step 1: Use TrinGalore for quality and adapter trimming of the raw RNA-seq reads, with followed codes:
trin_galore --quality 20 --phred33 --stringency 1 --length 20 --output_dir example/case1/output/trimmed --adapter file:example/case1/data/TruSeq-SE.fa example/case1/data/SRR1374921.fastq.gz
trin_galore --quality 20 --phred33 --stringency 1 --length 20 --output_dir example/case1/output/trimmed --adapter file:example/case1/data/TruSeq-SE.fa example/case1/data/SRR1374922.fastq.gz
trin_galore --quality 20 --phred33 --stringency 1 --length 20 --output_dir example/case1/output/trimmed --adapter file:example/case1/data/TruSeq-SE.fa example/case1/data/SRR1374923.fastq.gz
trin_galore --quality 20 --phred33 --stringency 1 --length 20 --output_dir example/case1/output/trimmed --adapter file:example/case1/data/TruSeq-SE.fa example/case1/data/SRR1374924.fastq.gz
Then, you finished the task. Step 2: Use HISAT2 for aligning the trimmed reads to the mouse m39 genome, with followed codes:
hisat2 -x example/case1/data/m39.fa -i example/case1/output/trimmed/SRR1374921.fastq.gz -S example/case1/output/SRR1374921.sam
hisat2 -x example/case1/data/m39.fa -i example/case1/output/trimmed/SRR1374922.fastq.gz -S example/case1/output/SRR1374922.sam
hisat2 -x example/case1/data/m39.fa -i example/case1/output/trimmed/SRR1374923.fastq.gz -S example/case1/output/SRR1374923.sam
hisat2 -x example/case1/data/m39.fa -i example/case1/output/trimmed/SRR1374924.fastq.gz -S example/case1/output/SRR1374924.sam
Do you want to modify codes? (y/n) (wait for 0.50 mins)
skip modify codes
[Round 2 finish]
```

人为中断、加载任意步骤状态，可以应对多种意外情况，比如：

1. 第n步报错了，但实际上根因错误在第m步。
2. 特定步骤任务卡死/崩溃/意外原因阻塞，可以手动中断当前步骤，并继承状态重试。
3. 用户想介入代码修改，但错过了修改时间窗口，也可以手动中断当前步骤，并继承状态重试。

## 2.15 开发环境 (Notebook)

### 2.15.1 Notebook 简介

EIHealth平台集成了基于开源的Jupyter Notebook，可为您提供在线的开发和调试工具，用于编写和调测模型训练代码。Notebook使您无需关心分析软件包的安装、升级和维护等工作，只需聚焦于科研工作，从而加快科研进展。

关于Jupyter Notebook的详细操作指导，请参见[Jupyter Notebook使用文档](#)。

## 支持的 AI 引擎

Notebook提供的AI引擎是Python 3，适配CPU/GPU芯片。

## 2.15.2 创建 Notebook

在开始进行模型开发前，您需要创建Notebook，并打开Notebook进行编码。

1. 在“项目管理 > 开发”页面，单击“创建Notebook”，参考表 [参数说明](#)填写信息。

图 2-230 创建 Notebook

The screenshot shows a web interface for creating a notebook. At the top, there's a breadcrumb '1234567890 / 创建Notebook'. The form has several sections:
 

- 名称**: A text input field containing 'notebook-test'.
- 描述**: A large text area for a description, with a character count '0/1024' at the bottom right.
- 镜像类型**: Two radio buttons, '系统' (selected) and '自定义'.
- 工作环境**: A dropdown menu showing 'PY3'.
- CPU**: A numeric input field with '1.0' and '+' '-' buttons, followed by '核'.
- GPU**: A numeric input field with '0.0' and '+' '-' buttons.
- 内存**: A numeric input field with '2.0' and '+' '-' buttons, followed by 'G'.
- 存储路径**: A text input field with a folder icon on the right.

 At the bottom, there are two buttons: '立即创建' (blue) and '取消' (grey).

表 2-34 参数说明

| 参数名称 | 说明                                                                                                                                                                                                                      |
|------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 名称   | Notebook的名称。名称长度1-63，仅支持小写字母、数字、中划线(-)，开始只能是小写字母，结束只能是小写字母或数字。                                                                                                                                                          |
| 描述   | Notebook的简要描述。                                                                                                                                                                                                          |
| 镜像类型 | 支持系统镜像和自定义镜像。                                                                                                                                                                                                           |
| 工作环境 | <ul style="list-style-type: none"> <li>● 选择系统镜像时，工作环境选择“PY3”。</li> <li>● 自定义镜像：当提供的镜像无法满足您的需求时，可以使用自己制作的Notebook镜像。镜像制作方法请参见<a href="#">镜像管理</a>。制作Notebook自定义镜像时，需要依赖平台提供的基础镜像，获取地址请参见<a href="#">获取镜像</a>。</li> </ul> |
| CPU  | 设置CPU大小。<br>取值范围：1~128，默认为1。                                                                                                                                                                                            |
| GPU  | 设置GPU大小。<br>取值范围：0~16，默认为0。                                                                                                                                                                                             |
| 内存   | 设置内存大小。<br>取值范围：2~512，默认为2。                                                                                                                                                                                             |

| 参数名称 | 说明                                                                                                                                                                                                                                                                       |
|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 存储路径 | <p>单击“存储路径”右侧  文件夹图标，设置用于存储 Notebook 数据的 OBS 路径。如果想直接使用已有的文件或数据，可将数据提前上传至对应的 OBS 路径下。</p> <p>包含本项目桶最多挂载 6 个，不包含本项目桶最多挂载 5 个。</p> <p>用户在 Notebook 列表的所有文件读写操作是基于所选择的 OBS 路径下的内容操作。</p> |

2. 参数填写无误后，单击“立即创建”，创建 Notebook。

## 2.15.3 Notebook 常用操作

### 打开 Notebook

1. 首次打开 Notebook 之前，请先[获取 url](#)，然后访问 url。  
然后单击“高级 > 继续前往 xxxx（不安全）”，页面显示“403: Forbidden”，然后返回 Notebook 列表页面。

#### 须知

请不要使用无痕浏览器访问。

2. 返回平台操作界面，在 Notebook 列表中，选择已经创建的 Notebook，单击“操作”列的“打开”，进入 Notebook 开发页面。

图 2-231 Notebook 开发页面



### 选择不同的 AI 引擎新建文件

1. 打开 Notebook 实例后，进入 Notebook 开发页面，在“File”页签下，单击文件夹名称进入子级。
2. 您可以单击右上角“New”，然后选择 AI 引擎“Python 3”，创建一个用于编码的文件。

#### 说明

Jupyter Notebook 页面不支持创建名称中带有特殊字符 < > ' " ; \ ` = # \$ % ^ & ( ) 的文件或文件夹。

图 2-232 选择 AI 引擎 “Python 3”



## 编写文件

新建文件后，自动进入文件编写页面，您也可以单击文件名称进入。

图 2-233 编写文件

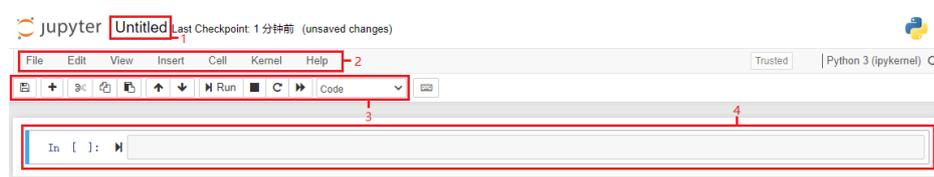
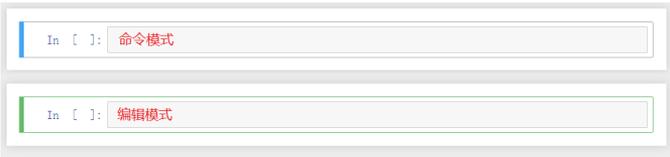


表 2-35 编写文件页面介绍

| 区域 | 区域名称 | 详细说明                                                                                                                                                                                                                                                                                                                                           |
|----|------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1  | 文件名称 | 您可以在此区域填写自定义的文件名称，修改保存后，对应的File列表也将发生变化。                                                                                                                                                                                                                                                                                                       |
| 2  | 菜单栏  | 菜单栏中有File、Edit、View、Insert、Cell、Kernel、Help等丰富功能。详细说明建议参考 <a href="#">Jupyter Notebook 使用文档</a> ，下方工具栏提供了常用的功能，能够满足常见的Python运行文件编写。                                                                                                                                                                                                            |
| 3  | 工具栏  | <p>工具栏罗列了支持的常用快捷操作，从左到右分别为：保存文件、添加新Cell、剪切选中的Cell、复制选中的Cell、粘贴选中的Cell、将选中Cell上移、将选中Cell下移、运行选中的Cell、终止kernel、重启kernel、重启kernel并重新运行所有Cell。</p> <p>Code下拉框中有四个选项：</p> <ul style="list-style-type: none"> <li>• Code：写Python代码</li> <li>• Markdown：写Markdown代码，通常用于注释</li> <li>• Raw NBConvert：转换工具</li> <li>• Heading：快捷添加Markdown标题</li> </ul> |

| 区域 | 区域名称   | 详细说明                                                                                                                                                                                                                                                  |
|----|--------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 4  | 代码Cell | <p>代码单元格。每一个Cell有两种模式：命令模式和编辑模式。</p> <p>最左侧为蓝色条时，是命令模式，绿色条表示编辑模式（此时Cell中有光标，可以进行代码编写）。在命令模式下，按下“Enter”键或者鼠标单击代码框可以进入编辑模式。在编辑模式下，按下“ESC”键或者鼠标单击代码框左侧区域即可进入命令模式。</p>  |

## 删除文件或文件夹

如果需要在Notebook中删除文件或文件夹，您可以在“Files”列表中勾选待删除的文件或文件夹，然后单击上方的红色删除按钮，即可删除选中的文件或文件夹。

### 说明

文件或文件夹删除完成后，需单击右上角的刷新按钮刷新页面，清除缓存文件。

图 2-234 删除

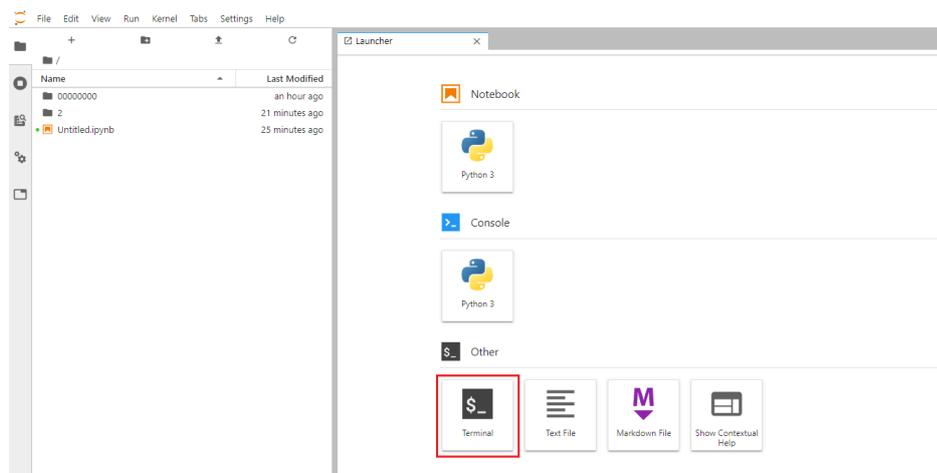


## 使用 Notebook Terminal 功能

对于习惯编码的开发者可以使用Terminal功能进行开发、调试和运行分析任务。

在“Files”页签下，单击右上角“Open JupyterLab”，然后选择“Terminal”，进入Terminal界面。

图 2-235 Terminal



例如，您可以执行wget命令在公开数据集中下载基因组测序数据。

图 2-236 执行命令



## 2.15.4 数据的上传和下载

在Notebook页面，可以通过“Upload”和“Download”上传和下载文件。上传和下载的文件大小限制为100MB。

当在Notebook中上传文件提示大小受限时，您可以根据以下不同场景将大文件上传或下载到Notebook中。

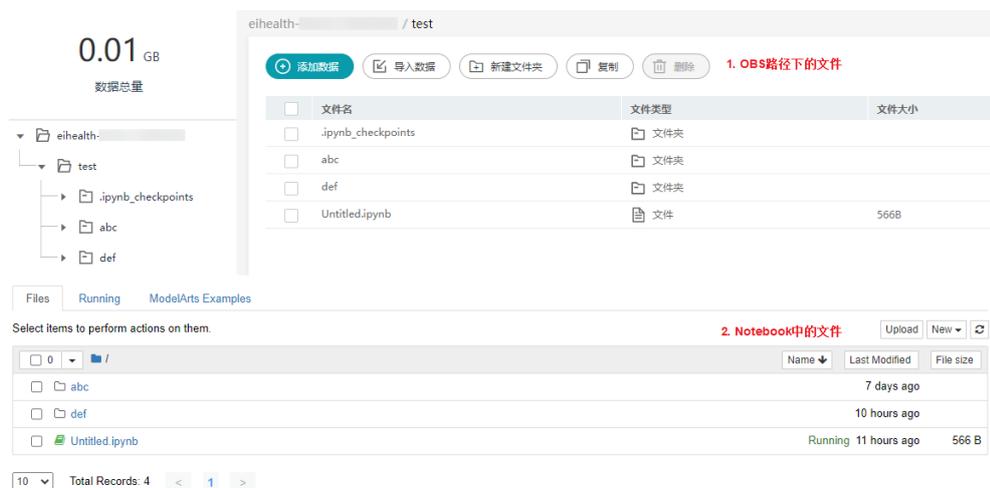
图 2-237 上传和下载文件



## OBS 存储类型的 Notebook

Notebook列表中所有文件的读写操作是基于所选择的OBS路径下的内容操作，即Notebook中的数据 and 被挂载的OBS路径中的数据是同步的。在OBS路径中创建文件夹、上传数据，会同步到Notebook中，Notebook中的操作也会同步到OBS中，如图通过OBS同步数据所示。

图 2-238 通过 OBS 同步数据



“Upload” 上传数据大小受限时，您可以通过以下多种方式将文件上传到OBS中，通过OBS与Notebook进行数据同步。

表 2-36 上传数据方法

| 上传方法      | 说明                                                                                                                   |
|-----------|----------------------------------------------------------------------------------------------------------------------|
| “数据”页面上传  | 通过“数据”页面上传数据，支持上传最大为1GB的单个文件。<br>数据上传方法请参见 <a href="#">“数据”页面上传</a> 。                                                |
| 使用命令行工具上传 | 命令行工具（eihealth-toolkit）配套EIHealth平台，提供数据、应用、流程和作业资源的管理和使用，支持上传最大为48.8TB的单个文件。<br>数据上传方法请参见 <a href="#">命令行工具概述</a> 。 |

### 📖 说明

对于非挂载目录以外的目录下的文件，重启Notebook后会消失。例如，上传文件至Notebook的根目录下，该文件并不在被挂载的obs路径中，重启Notebook，该文件会消失。

## 2.15.5 自定义镜像创建 Notebook 样例

Notebook提供了在线的开发和调试工具，用于编写和调测模型训练代码，在创建Notebook时，平台为您提供了系统镜像。同时，您也可以使用自定义镜像，创建Notebook，满足个性化的开发需求。

本章节介绍如何使用自定义镜像创建Notebook（在基础镜像中安装化学分子格式转换工具Open Babel），详细步骤如下所示：

- [步骤1：安装容器引擎](#)
- [步骤2：获取Notebook基础镜像](#)
- [步骤3：制作并上传镜像](#)

- **步骤4: 创建并使用Notebook**

## 步骤 1: 安装容器引擎

在制作自定义镜像时，您需要准备一台安装有Docker的机器。请使用自己的电脑搭建Docker环境，或者使用华为云弹性云服务器ECS搭建Docker环境。

例如，在Linux操作系统下，可以使用如下命令快速安装容器引擎。

```
curl -fsSL get.docker.com -o get-docker.sh
sh get-docker.sh
```

## 步骤 2: 获取 Notebook 基础镜像

创建Notebook时所需的自定义镜像，依赖于医疗智能体平台自研的基础镜像，您需要基于获取的基础镜像制作自定义镜像。

先连接容器镜像服务，参考[步骤1.连接容器镜像服务](#)操作，然后使用如下的镜像地址拉取基础镜像。

```
# 基础镜像
docker pull swr.${region}.myhuaweicloud.com/eihealth-notebook/eihealth-notebook:${image_version}
```

### 📖 说明

- `${region}`: 需根据用户实际购买region填写。例如，`${region}`为cn-north-4。
- `${image_version}`: 需根据实际版本填写。例如，`${image_version}`为cuda11.0-custom-v1.0.8。

在基础镜像中，为您内置了CUDA11.0环境。使用CUDA11.0环境前，需要执行以下命令导入环境变量。

```
export PATH=$PATH:/usr/local/nvidia/bin
export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:/usr/local/nvidia/lib64
```

## 步骤 3: 制作并上传镜像

### 1. 制作镜像。

- 【可选】根据网络情况，配置基础镜像中的PyPi Mirror，对下载进行加速。基础镜像中的PyPi Mirror，默认配置为华为云软件开发云的PyPi mirror。您可以在容器中执行如下命令，查看PyPi Mirror。如果您想用其他PyPi Mirror，可将命令中的index-url参数修改为您需要的PyPi mirror。

```
cat /root/.pip/pip.conf.product
[global]
index-url = http://repo.myhuaweicloud.com/repository/pypi/simple
format = columns
[install]
trusted-host=repo.myhuaweicloud.com
```

- 安装所需软件。本例中使用cpu基础镜像，并安装化学分子格式转换工具Open Babel。

```
# 创建并进入Dockerfile文件
vi Dockerfile
# 编写Dockerfile文件，安装Open Babel
FROM swr.cn-north-4.myhuaweicloud.com/eihealth-notebook/eihealth-notebook:cuda11.0-custom-v1.0.8
RUN sudo apt-get update
RUN sudo apt-get install -y openbabel
```

### 说明

- Dockerfile中不可以指定CMD以及ENTRYPOINT，否则会覆盖基础镜像启动脚本，引起异常。
- c. 按键盘Esc键，并执行:wq保存退出Dockerfile。
  - d. 执行**docker build -t image:tag .**命令，自动完成镜像制作。  
命令中image为镜像名称，tag为镜像标签，名称可自定义。  
详细镜像制作过程请参见[制作Docker镜像](#)和[Dockerfile参考](#)。
2. 上传镜像。
    - a. [上传镜像](#)。
    - b. 在“项目管理”页面“镜像”页签中，可查看已上传的镜像。
    - c. 在平台镜像管理列表中，将已上传的镜像分类为“NOTEBOOK”。

## 步骤 4：创建并使用 Notebook

1. 在医疗智能体平台选择“项目管理 > 开发”，进入Notebook管理页面。

图 2-239 创建 Notebook



2. 单击“创建Notebook”，并参考如下的参数说明进行填写。

表 2-37 参数说明

| 参数名称 | 说明                                                             |
|------|----------------------------------------------------------------|
| 名称   | Notebook的名称。名称长度1-63，仅支持小写字母、数字、中划线(-)，开始只能是小写字母，结束只能是小写字母或数字。 |
| 描述   | Notebook的简要描述。                                                 |
| 镜像类型 | 选择自定义镜像。                                                       |
| 工作环境 | 选择已上传的自定义镜像。                                                   |
| CPU  | 设置CPU大小。                                                       |
| GPU  | 设置GPU大小。                                                       |
| 内存   | 设置内存大小。                                                        |

| 参数名称 | 说明                                                                                                                                                                                                                                |
|------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 存储路径 | <p>单击“存储路径”右侧  文件夹图标，设置用于存储 Notebook 数据的 OBS 路径。如果想直接使用已有的文件或数据，可将数据提前上传至对应的 OBS 路径下。</p> <p>用户在 Notebook 列表的所有文件读写操作是基于所选择的 OBS 路径下的内容操作。</p> |

3. 参数确认无误后，单击“立即创建”，完成 Notebook 的创建操作。
4. 在 Notebook 列表中，当“状态”由“启动中”转变为“运行中”后，单击操作列的“打开”。

您可以直接使用该 Notebook，编写和调测模型，进行开发工作。关于 Jupyter Notebook 的详细操作指导，请参见 [Jupyter Notebook 使用文档](#)。

图 2-240 Notebook



## 2.15.6 JupyterLab 简介及常用操作

JupyterLab 是一个交互式的开发环境，是 Jupyter Notebook 的下一代产品，可以使用它编写 Notebook、操作终端、编辑 Markdown 文本、打开交互模式、查看 csv 文件及图片等功能。

可以说，JupyterLab 是开发者们下一阶段更主流的开发环境。JupyterLab 支持更加灵活和更加强大的项目操作方式，但具有和 Jupyter Notebooks 一样的组件。

### 打开 JupyterLab

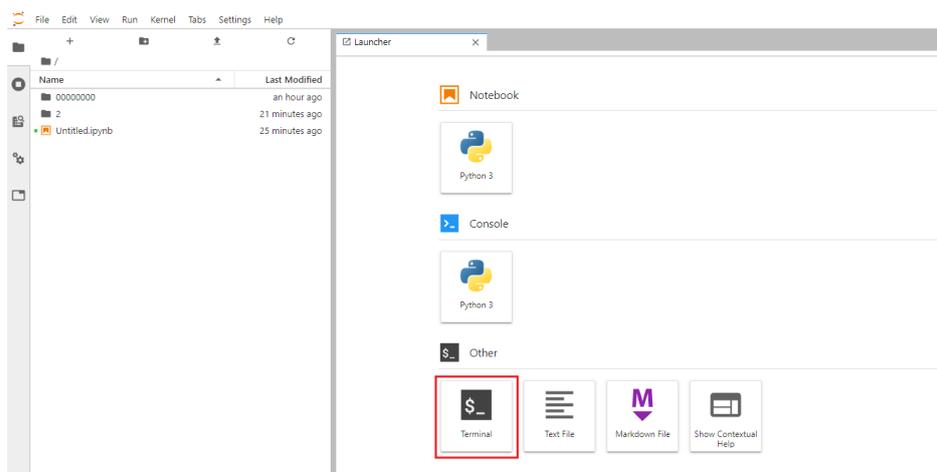
1. 选择状态为“运行中”的 Notebook 实例，单击“操作”列的“打开”访问 Notebook。
2. 单击右上角的“Open JupyterLab”，可直接打开此 Notebook 实例对应的 JupyterLab 页面。

图 2-241 进入 JupyterLab



3. 进入 JupyterLab 页面后，自动打开 Launcher 页面，如下图所示。您可以使用开源支持的所有功能，详细操作指导可参见 [JupyterLab 官网文档](#)。

图 2-242 JupyterLab 主页

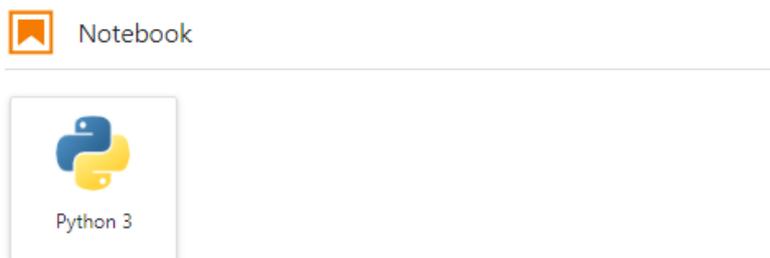


## 新建并打开 Notebook

进入JupyterLab主页后，可在“Notebook”区域下，选择适用的AI引擎，单击后将新建一个对应框架的Notebook文件。

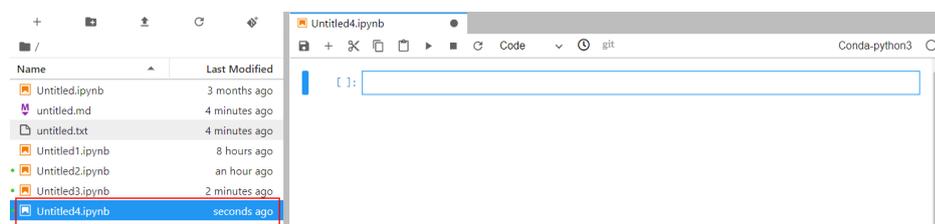
由于每个Notebook实例选择的工作环境不同，其支持的AI框架也不同，下图仅为示例，请根据实际显示界面选择AI框架。

图 2-243 选择 AI 引擎并新建一个 Notebook



新建的Notebook文件将呈现在左侧菜单栏中。

图 2-244 新建文件



## 新建文件并打开 Console

Console的本质为Python终端，输入一条语句就会给出相应的输出，类似于Python原生的IDE。

进入JupyterLab主页后，可在“Notebook”区域下，选择适用的AI引擎，单击后将新建一个对应框架的Notebook文件。

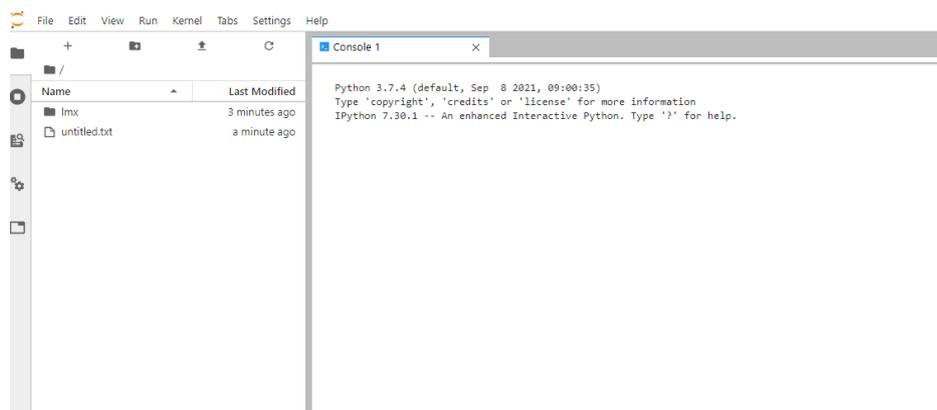
由于每个Notebook实例选择的工作环境不同，其支持的AI框架也不同，下图仅为示例，请根据实际显示界面选择AI框架。

图 2-245 选择 AI 引擎并新建一个 Console



文件创建成功后，将直接呈现Console页面。

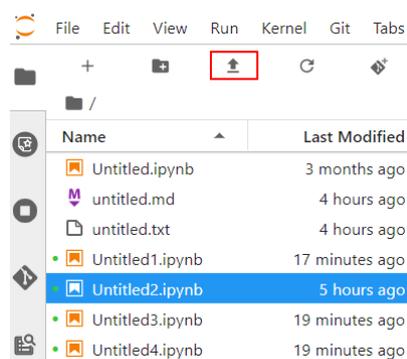
图 2-246 新建文件（ Console ）



## 上传文件

进入JupyterLab页面后，您可以单击左上角“Upload”快捷键，从本地选择一个文件上传。

图 2-247 上传文件

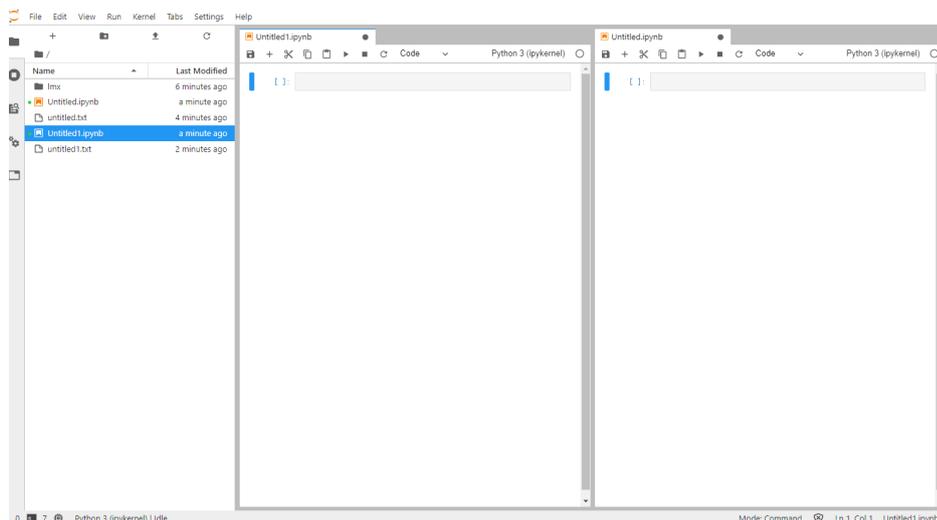


## 编辑文件

JupyterLab可以在同一个窗口同时打开几个Notebook或文件（如HTML、TXT、Markdown等），以页签形式展示。

JupyterLab的一大优点是，可以任意排版多个文件。在右侧文件展示区，您可以拖动打开文件，随意调整文件展示位置，可以同时打开多个文件。

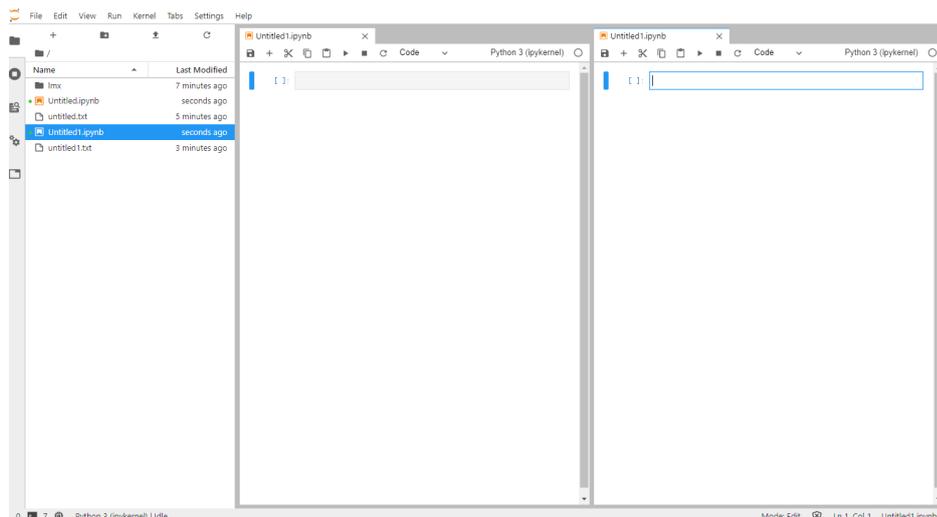
图 2-248 多文件任意编排



当在一个Notebook中写代码时，如果需要实时同步编辑文件并查看执行结果，可以新建该文件的多个视图。

打开此文件，然后单击菜单栏“File>New View for Notebook”，即可打开多个视图。

图 2-249 同一个文件的多个视图

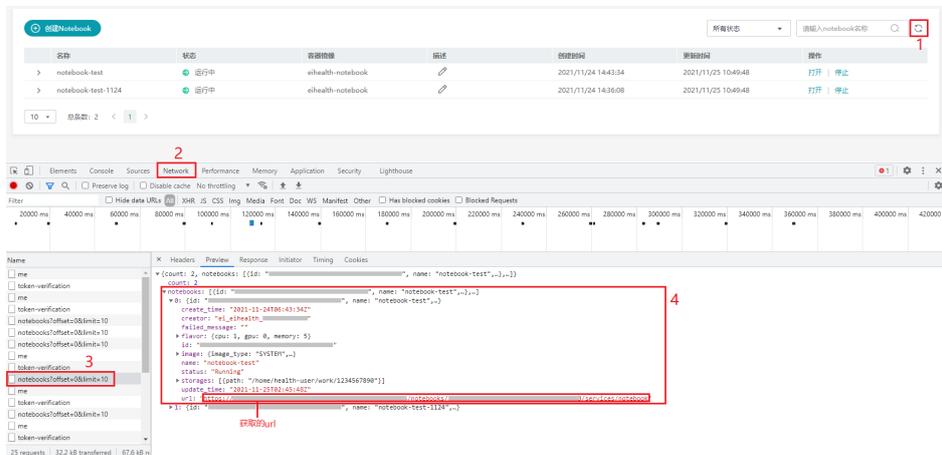


### 2.15.7 url 获取方法

1. 在Notebook列表页面，按F12键打开，切换至Network页签。

- 单击页面右侧  按钮，在Network页签的“Name”列单击“notebooks?offset=0&limit=10”，右侧展开代码行，在“url”行获取url链接，如下图所示。

图 2-250 获取 url



## 2.15.8 Notebook 安装 Conda 指导

- 打开Notebook，在“File”页签下选择“Terminal”。

图 2-251 选择 Terminal



- 下载和安装Anaconda。
  - 获取Repository和Anaconda安装包。  
Repository: <https://repo.anaconda.com/archive/index.html>  
Anaconda Installer: [https://repo.anaconda.com/archive/Anaconda3-2022.10-Linux-x86\\_64.sh](https://repo.anaconda.com/archive/Anaconda3-2022.10-Linux-x86_64.sh)
  - 安装Anaconda。
    - 执行如下命令。  

```
wget https://repo.anaconda.com/archive/Anaconda3-2022.10-Linux-x86_64.sh --no-check-certificate
bash Anaconda3-2022.10-Linux-x86_64.sh
```
    - 在第一个暂停中按“Enter”继续。  

```
[health-user@notebook-49129a4a-7b2e-4631-acc5-dcbb479c1093-186864c6-tp49v tmp]$ bash Anaconda3-2022.10-Linux-x86_64.sh
Welcome to Anaconda3 2022.10
In order to continue the installation process, please review the license
agreements.
Please, press ENTER to continue
>>>
```
    - 在第二个暂停中，将显示用户许可协议，按“Enter”阅读更多信息。

```

End User License Agreement - Anaconda Distribution
-----
Copyright 2015-2022, Anaconda, Inc.

All rights reserved under the 3-clause BSD License:

This End User License Agreement (the "Agreement") is a legal agreement between you and Anaconda, Inc. ("Anaconda") and governs your use of Anaconda Distribution (which was formerly known as Anaconda Individual Edition).

Subject to the terms of this Agreement, Anaconda hereby grants you a non-exclusive, non-transferable license to:

* Install and use the Anaconda Distribution (which was formerly known as Anaconda Individual Edition),
* Modify and create derivative works of sample source code delivered in Anaconda Distribution from Anaconda's repository, and;
* Redistribute code files in source (if provided to you by Anaconda as source) and binary forms, with or without modification subject to the requirements set forth below, and;

Anaconda may, at its option, make available patches, workarounds or other updates to Anaconda Distribution. Unless the updates are provided with their separate governing terms, they are deemed part of Anaconda Distribution licensed to you as provided in this Agreement. This Agreement does not entitle you to any support for Anaconda Distribution.

Anaconda reserves all rights not expressly granted to you in this Agreement.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:
    
```

iv. 阅读完成后，输入“yes”，然后按“Enter”。

```

You must comply with all domestic and international export laws and regulations that apply to the software, which include restrictions on destinations, end users, and end use. Anaconda Distribution includes cryptographic software. The country in which you currently reside may have restrictions on the import, possession, use, and/or re-export to another country, of encryption software. BEFORE using any encryption software, please check your country's laws, regulations and policies concerning the import, possession, or use, and re-export of encryption software, to see if this is permitted. See the Wassenaar Arrangement http://www.wassenaar.org/ for more information.

Anaconda has self-classified this software as Export Commodity Control Number (ECCN) 1A999 which includes mass market information security software using or performing cryptographic functions with asymmetric algorithms. No license is required for export of this software to non-embargoed countries.

The Intel Math Kernel Library contained in Anaconda Distribution is classified by Intel as ECCN 5D992.c with no license required for export to non-embargoed countries.

The following packages listed on https://www.anaconda.com/cryptography are included in the repository accessible through Anaconda Distribution that relate to cryptography.

Last updated February 25, 2022

Do you accept the license terms? [yes/no]
[no] >>>
Please answer 'yes' or 'no':
>>>
    
```

v. 选择安装位置，按“Enter”。

```

Please answer 'yes' or 'no':
>>> yes

Anaconda3 will now be installed into this location:
/home/health-user/anaconda3

- Press ENTER to confirm the location
- Press CTRL-C to abort the installation
- Or specify a different location below

[/home/health-user/anaconda3] >>>
    
```

vi. 安装完成后，输入“yes”。

```

done
installation finished.

WARNING:
You currently have a PYTHONPATH environment variable set. This may cause
unexpected behavior when running the Python interpreter in Anaconda3.
For best results, please verify that your PYTHONPATH only points to
directories of packages that are compatible with the Python interpreter
in Anaconda3: /home/health-user/anaconda3
Do you wish the installer to initialize Anaconda3
by running conda init? [yes/no]
[no] >>>
    
```

3. 初始化Anaconda配置并测试安装。

a. 执行以下命令初始化Anaconda配置。

```

export PATH=~/.anaconda3/bin:$PATH
source ~/.bashrc
    
```

b. 测试安装。

```

conda -V
    
```

4. 查看当前环境列表，并创建新环境参数。

```

conda env list
conda create -n py37 python=3.7
    
```

5. 输入“y”，按“Enter”继续。

```

The following NEW packages will be INSTALLED:

 libgcc_mtex pkgs/main/linux-64::libgcc_mtex-0.1-main None
 _openmp_mutex pkgs/main/linux-64::_openmp_mutex-5.1-l_gnu None
 ca-certificates pkgs/main/linux-64::ca-certificates-2022.10.11-h06a4308_0 None
 certifi pkgs/main/linux-64::certifi-2022.9.24-py37h06a4308_0 None
 ld_impl_linux-64 pkgs/main/linux-64::ld_impl_linux-64-2.38-h1181459_1 None
 libffi pkgs/main/linux-64::libffi-3.4.2-h6a67895_6 None
 libgcc-ng pkgs/main/linux-64::libgcc-ng-11.2.0-h1234567_1 None
 libgomp pkgs/main/linux-64::libgomp-11.2.0-h1234567_1 None
 libstdcxx-ng pkgs/main/linux-64::libstdcxx-ng-11.2.0-h1234567_1 None
 ncurses pkgs/main/linux-64::ncurses-6.3-h5ee18b_3 None
 openssl pkgs/main/linux-64::openssl-1.1.1a-h7f9727e_0 None
 pip pkgs/main/linux-64::pip-22.3.1-py37h06a4308_0 None
 python pkgs/main/linux-64::python-3.7.15-h7a1cb2a_1 None
 readline pkgs/main/linux-64::readline-8.2-h5ee18b_0 None
 setuptools pkgs/main/linux-64::setuptools-65.5.0-py37h06a4308_0 None
 sqlite pkgs/main/linux-64::sqlite-3.40.0-h598229e_0 None
 tk pkgs/main/linux-64::tk-8.6.12-h1ccaba5_0 None
 wheel pkgs/main/noarch::wheel-0.37.1-pyhd3eb1b0_0 None
 xz pkgs/main/linux-64::xz-5.2.8-h5ee18b_0 None
 zlib pkgs/main/linux-64::zlib-1.2.13-h5ee18b_0 None

Proceed ([y]/n)? y
    
```

6. 切换到新环境。

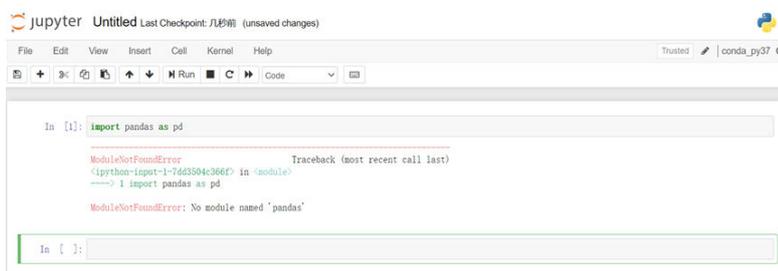
```

conda activate py37
    
```

7. 生成Notebook kernel。  
conda install -n py37 ipykernel  
python -m ipykernel install --user --name py37 --display-name "conda\_py37"
8. 使用新生成的Notebook kernel。
  - a. 在Files页面，选择新建的Notebook kernel。



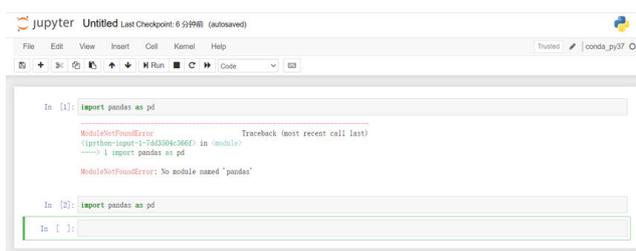
- b. 安装新软件包。
    - i. 查看pandas是否安装。如下表示 pandas未安装。



- ii. 返回Terminal，切换到相应的环境“py37”，然后安装pandas。  
conda activate py37  
conda install -n py37 pandas



- iii. 安装完成后，返回Notebook,然后使用新软件包。



## 参考信息

如何修复“conda: command not found”？

当terminal第二次打开时，环境变量将被清理，因此我们需要再次初始化Anaconda配置。

```
export PATH=~/.anaconda3/bin:$PATH
source ~/.bashrc
```

## 2.16 大规模药物虚拟筛选

### 2.16.1 “神农项目”简介

“神农项目”是完全免费公开的新冠药物虚拟筛选数据库。在抗疫期间，为了寻找有效治愈新冠肺炎的治疗方式，华为云医疗智能体项目团队联合多家科研机构，从新冠病毒蛋白序列开始，针对所有21个靶点蛋白进行同源建模、分子动力学模拟优化，获取了靶点蛋白的3D结构。并对超过8500个已上市、进入临床的小分子药物进行了约18万种药物-靶点配对情况的计算评估，让研究人员可以同时从21个蛋白的角度，综合、无偏地评估药物效果，从而为后续的药物机制研究、临床试验提供线索。同时，该研究成果以封面杂志发表在化学信息学领域国际顶级期刊《Journal of Chemical Information and Modeling》(化学信息与建模) 2020年12月新冠特刊中。

论文链接: <https://doi.org/10.1021/acs.jcim.0c00821>

### 查看“神农项目”药物筛选结果

登录EIHealth平台，可在“专题”页签中查看“神农项目”药物筛选结果。可参考[视频教程](#)了解详细的使用指导。

在图2-253中，页面最右侧显示结合能的最大值和最小值，以及对应的列名称、行名称。

单击列名称或行名称，在弹窗中单击图标，可以对结合能数据进行排序。

单击图标，在弹窗中可以自定义选择要显示的列。

图 2-252 神农项目

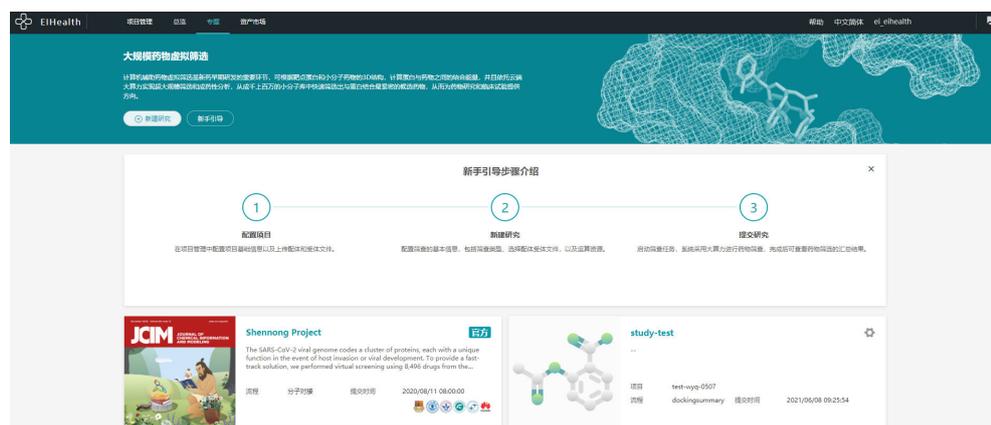
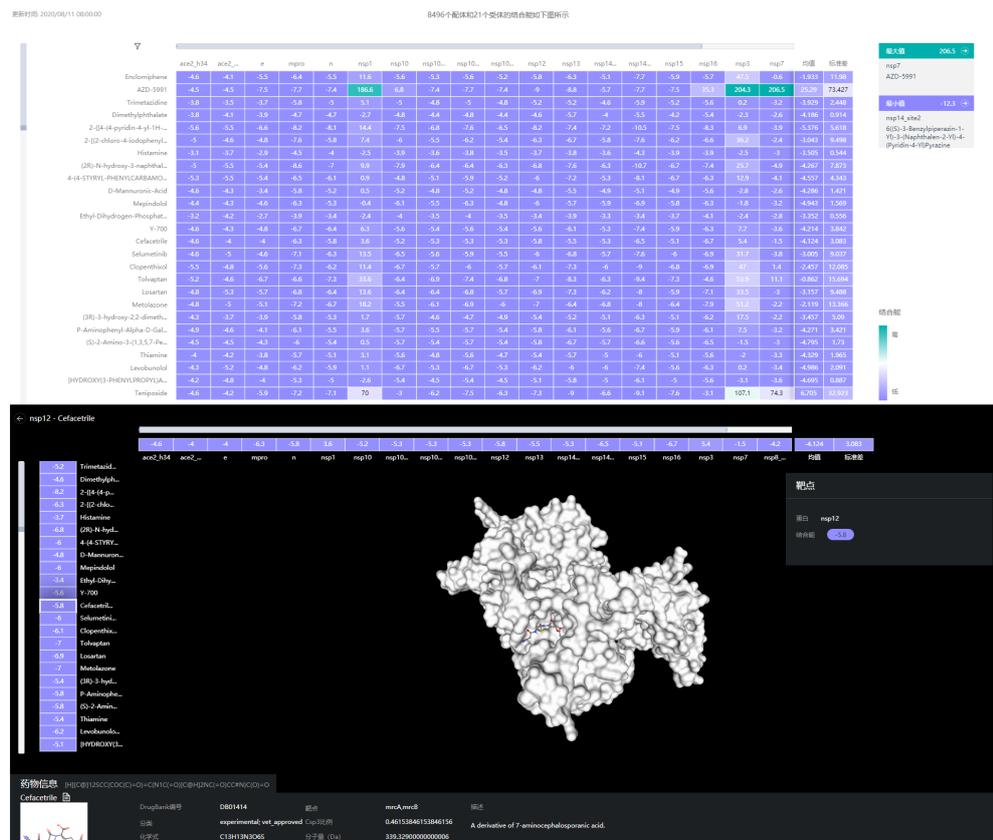


图 2-253 药物筛选结果



## 2.16.2 药物虚拟筛选

计算机辅助药物虚拟筛选是新药早期研发的重要环节，可根据靶点蛋白和小分子药物的3D结构，计算蛋白与药物之间的结合能量，并且依托云端大算力实现超大规模筛选和成药性分析，从成千上万的小分子库中快速筛选出与蛋白结合最紧密的候选药物，从而为药物研究和临床试验提供方向。药物虚拟筛选以项目成员为粒度进行资源的隔离和访问控制，不同成员角色对数据库的操作权限请参见[成员角色和权限](#)。

### 创建药物虚拟筛选任务

虚拟药物筛选支持使用资产市场中预置的“Docking Summary”流程对小分子化合物配体和蛋白受体进行对接。该流程可以实现以下功能。

- 整合分子对接结果，生成结合能矩阵。
- 整合受体与分子对接产生的配体构象，用于可视化展示。
- 对配体分子进行注释，包括DrugBank编号、分类、化学式、XLOGP3、TPSA、靶点、Csp3比例、分子量、可旋转键数目。

使用步骤如下所示。

1. 在“资产市场”中订阅版本为1.0.0“Docking Summary”流程至所需的项目中。
2. 进入“专题”页签，单击“新建研究”。
3. 填写任务的基本信息，包括选择任务所属项目，研究的名称和描述。

图 2-254 基本信息

4. 选择配体分子和受体蛋白。
  - 作业名称：自定义名称。
  - 类型：选择小分子化合物。
  - 流程：选择从资产市场中订阅的“Docking Summary”流程。
    - 配体分子：配体分子文件，支持SMILES、3D SDF、PDB、MOL2格式。
  - 受体蛋白：受体蛋白文件，支持PDB格式。

图 2-255 选择配体分子和受体蛋白

选择配体分子和受体蛋白

5. 设置数据库。
 

数据库功能可以将任务运行过程中产生的数据文件按照模板生成数据库。

  - 数据库名称：数据库的名称。
  - 输出文件格式：可以将流程生成的分子对接结果，保存为.txt、.csv或.vcf格式。使用“Docking Summary”流程时，保存格式为.txt。
  - 相对路径：流程运行完成后，会按照流程子任务的名称生成数据文件，相对路径指按照哪个数据路径中的结果文件生成数据库。
 

对于“Docking Summary”流程，包含5个子任务，默认在task-5-docking summary中保存有汇总的数据文件。

    - task-1-ligand 3dsdf to pdbqt：将配体的sdf文件转换为pdbqt文件。
    - task-2-ligand smiles to 3dsdf：将配体的smiles文件转换为3dsdf文件。
    - task-3-receptor pdb to pdbqt：将受体的pdb文件转换为pdbqt文件。
    - task-4-qvina-w：分子对接。
    - task-5-docking summary：汇总分子对接结果。

图 2-256 数据路径和流程图

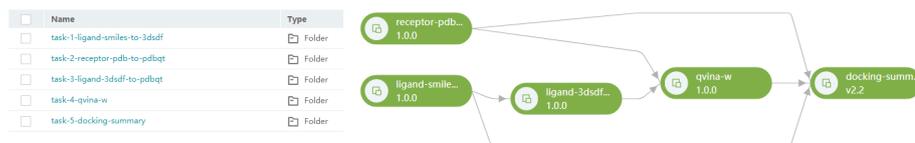


图 2-257 设置数据库

数据库

\* 数据库名称

\* 文件格式 \*.txt \*.csv \*.vcf

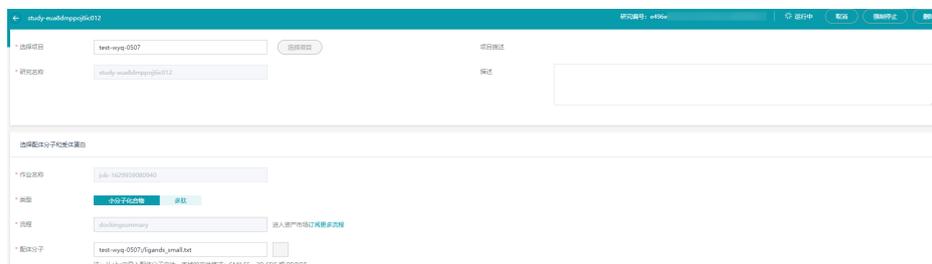
\* 文件路径

6. 设置完成后，单击“提交”，执行药物虚拟筛选任务。

对于“运行中”的任务，单击 图标，允许取消、强制停止或删除。

对于“已取消”、“运行失败”的任务，单击 图标，允许修改任务参数，再次提交任务。

图 2-258 运行状态



## 查看执行结果

药物和蛋白结合能通过热点图呈现，并统计出结合能的均值和标准差。单击热点图，可查看详细3D分子对接结果、结合能排序、药物注释信息。

在图2-259中，页面最右侧显示结合能的最大值和最小值，以及对应的列名称、行名称。

单击列名称或行名称，在弹窗中单击 图标，可以对结合能数据进行排序。

单击 图标，在弹窗中可以自定义选择要显示的列。

图 2-259 结合能图

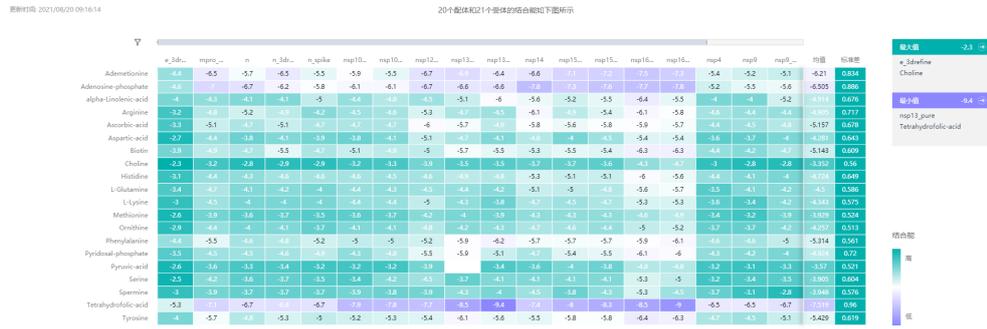


图 2-260 对接结果

